AD-A061 858 PENNSYLVANIA STATE UNIV UNIVERSITY PARK APPLIED RESE--ETC F/G 17/2
THE IDENTIFIABILITY OF APPROXIMATE VOCAL TRACT IMPULSE RESPONSE--ETC(U)
DEC 77 F S MCKENDREE N00017-73-C-1418
UNCLASSIFIED ARL/PSU/TM-77-331 NL

1 OF 2
AD A061858

(12) LEVEL II

(6) THE IDENTIFIABILITY OF APPROXIMATE VOCAL TRACT IMPULSE RESPONSE MAGNITUDES

(10) Francis Speed McKendree

(11) 16 Dec 77

(9) Master's thesis,

(12) 192 p.

Technical Memorandum
File No. TM 77-331
December 16, 1977
Contract No. (15) N00017-73-C-1418

Copy No. 6

(14) ARL/PSU/TM-77-331

The Pennsylvania State University
Institute for Science and Engineering
APPLIED RESEARCH LABORATORY
Post Office Box 30
State College, PA 16801

NAVY DEPARTMENT

NAVAL SEA SYSTEMS COMMAND

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER TM 77-331	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) THE IDENTIFIABILITY OF APPROXIMATE VOCAL TRACT IMPULSE RESPONSE MAGNITUDES		5. TYPE OF REPORT & PERIOD COVERED M.S. Thesis, November 1978
		6. PERFORMING ORG. REPORT NUMBER TM 77-331
7. AUTHOR(s) Francis Speed McKendree		8. CONTRACT OR GRANT NUMBER(s) N00017-73-C-1418
9. PERFORMING ORGANIZATION NAME AND ADDRESS The Pennsylvania State University Applied Research Laboratory P. O. Box 30, State College, PA 16801		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Sea Systems Command Department of the Navy Washington, DC 20362		12. REPORT DATE December 16, 1977 13. NUMBER OF PAGES 100 pages & figures
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified, Unlimited 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited,
per NSSC (Naval Sea Systems Command), 1/30/78

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

speech signal
speech analysis
vocal impulse

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

It is shown that the vocal tract impulse response magnitude should be less variable for a given speaker than other acoustic measures of his speech. Cepstrum analysis is used to deconvolve the vocal tract impulse response and glottal pressure wave of each of 1850 speech segments taken from running English speech. Linear correlation coefficients derived from pairs of impulse responses are shown to differ, depending upon whether the two impulse responses were taken from the same speaker's utterances, from speakers of the same sex and/or vocal history, or from altogether different speakers. A feasible method of speaker identification

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT (Continued)

which can in principle operate automatically is developed from this approach and tested.

ABSTRACT

It is shown that the vocal tract impulse response magnitude should be less variable for a given speaker than other acoustic measures of his speech. Cepstrum analysis is used to deconvolve the vocal tract impulse response and the glottal pressure wave of each of 1850 speech segments taken from running English speech. Linear correlation coefficients derived from pairs of impulse responses are shown to differ, depending upon whether the two impulse responses were taken from the same speaker's utterances, from speakers of the same sex and/or vocal history, or from altogether different speakers. A feasible method of speaker identification which can in principle operate automatically is developed from this approach and tested.

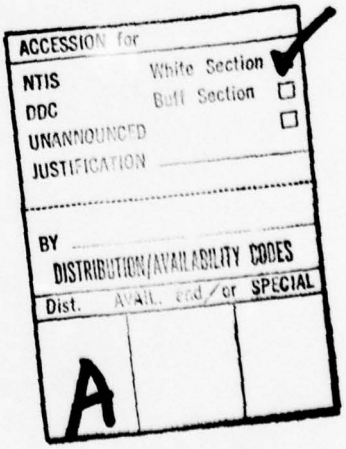

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont.)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (cont.)

LIST OF SYMBOLS AND NOTATION

f_0	voice fundamental frequency
f_1	first formant frequency
f_n	n^{th} formant frequency
$F_f( )$	Fourier transform operator $\int_{-\infty}^{\infty} ( )e^{-jft}dt$ (direct operator)
$F_t( )$	Fourier transform operator $\int_{-\infty}^{\infty} ( )e^{jft}df$ (inverse operator)
s(t)	measured time function
g(t)	glottal pressure wave time function
h(t)	vocal tract impulse response
S(f)	direct Fourier transform of s(t) , or measured spectrum
G(f)	glottal pressure wave spectrum
H(f)	vocal tract frequency response
| |	complex magnitude operator, $\|x+jy\| = \sqrt{x^2 + y^2}$
/ /	denotes spelling in phonetic symbols within the text body
CV	schematic for an utterance in which a consonant precedes a vowel
VC	schematic for an utterance in which a vowel precedes a consonant
VCV	schematic for a vowel-consonant-vowel utterance
p_t	probability of a true decision
p_f	probability of a false decision
$p_t(n)$	probability that n trials will give a true result
$p_f(n)$	probability that n trials will give a false result
$p_i(n)$	probability that n trials will give an indeterminate result
R(n)	reliability of a classification or identification task

LIST OF PHONETIC SYMBOLS

STANDARD SYMBOL	APPROXIMATE SOUND	NUMBER* EQUIVALENT
/ɪ/	bit	2
/ɛ/	fed	4
/æ/	sad	5
/ɔ/	awed	8
/ʌ/	bud	9
/ɛi/	take	32**
/ɔi/	type	102**

* The number equivalent provides a code which is intelligible to the computer and to the system operator.

** Compound symbol describing a diphthong used in this thesis.

ACKNOWLEDGMENTS

The author is very grateful for the patience, help, and encouragement of Dr. Harvey R. Gilbert and Prof. Lawrence C. Pharo, members of his thesis committee, and of Prof. Richard O. Rowlands, his thesis adviser:

If this work pleases, theirs is the credit:

If not, mine is the blame.

Many persons were of great help in this work. Among them are Mr. Claus P. Janota, who helped the author in the use of both hardware and software, and who brought the mathematical technique of cepstrum analysis to the author's attention; Mr. Thomas B. Way and Mr. Robert F. Hoover of the Noise Analysis Room staff; and Mr. Leonard L. Holliday and Mr. Jack W. Sharer of the ARL computer staff.

CHAPTER I

INTRODUCTION

1.1 History

Several methods have been developed for recording the amplitude versus frequency distribution of the speech signal as a function of time. The distinctive appearance of such a display leads to the conjecture that such displays might form the basis of a method of speaker identification.

Early researchers used several laborious methods to measure the spectra of successive parts of a speech signal. In one experiment, segments of a sentence were graphed as a function of time and subjected to manual Fourier analysis. The resulting spectra were combined into a "three-dimensional" graph with time as the abscissa and frequency as the ordinate; the relative intensity was indicated by the degree of darkness of the various parts of the display [1]. In another experiment, a group of ten bandpass filters was used to separate a sentence into components within each of the frequency bands. The output of each filter was recorded for subsequent study [2].

In 1946, Koenig, Dunn, and Lacey, working at the Bell Telephone Laboratories, reported the development of a "sound spectrograph" [3]. In this device, a signal of 2.4 seconds duration was recorded on a rotating drum which was in turn geared to a wave analyzer and a

recording device using electrosensitive paper. The compressed output of the wave analyzer was used to control the darkness of the trace on the paper. In this way, a vertical section of the trace represented the instantaneous spectrum of the corresponding part of the signal which was recorded on the drum below it. The original device was sufficiently rapid in response to be able to record the spectrum of each glottal wave pulse. A series of commercial devices embodying this principle is presently marketed by Kay-Bee Elemetrics under the generic name of sonagraph. The display produced by the device is termed a sonagram.

Figure 1.1 is a sonagram of the author saying, "Joe took father's shoe bench out." At the left end of the display is a calibration signal which is generated by the sonagraph. It consists of a 500-Hertz ramp wave, and the harmonics of the ramp appear as dark bars in the display.

The speech portion of the display clearly shows the distinction between the voiced and unvoiced signals. The unvoiced segments have broad-band frictional noise which is filtered by the vocal tract, as is indicated by the varying darkness of "f" in father, "sh" in shoe, and "ch" in bench. The voiced segments show vertical striations, each of which displays the spectrum of a single cycle of the signal. The "j" of Joe is a voiced consonant, and therefore shows both broad-band frictional noise and the vertical striations characteristic of voiced sounds.

The dark bars in the display indicate the frequency bands of maximum energy. These are termed formants, and the frequencies at which they lie are called the formant frequencies. The first (i.e., the lowest frequency) formant will be denoted by f_1, the second by f_2, and so on. The voice fundamental frequency will be symbolized by f_0.

The formant structure is the primary clue used by the auditory system in identifying a sound, whether it is a speech segment, a musical tone, or any other noise. The formant structure is characteristic of the sound; compare, for example, the "oo" of took, the "o" of shoe, and the last part of the "ou" of out. Each of these is very similar in appearance, and similar in sound to the ear.

Speaker identification experiments may be grouped for purposes of discussion in the following way. We first distinguish between "sorting" and "identification" tasks. In a sorting task, it is desired to determine which of a set of samples of different speakers was uttered by each of the speakers. In an identification task, it is desired to determine which of a set of known speakers produced a given unknown utterance. One next distinguishes between open and closed tasks. In a closed task, all samples required to complete the task are known to be included in the set of samples. In an open task, it is possible that some of the required samples are not in the set.

L. G. Kersta, in 1962 [4], reported the success of a method of speaker identification utilizing a voice spectrograph. The technique was dubbed "voiceprinting," by analogy with the established identification technique of fingerprinting. Kersta reported that female high school sophomores, after one week of instruction and practice, were highly successful in identifying speakers on the basis of their spectrograms of common English words. In a closed sorting experiment with five to twelve speakers, error rates of 0.8 percent were obtained for words taken in isolation, and 1.0 percent for words taken from context in running speech.

Other researchers worked subsequently to verify and to extend these results. Tosi, et al., performed a similar sorting task [5]. Their results were very like those reported by Kersta, but they note in their conclusions ". . . that this group of trials does not fit any type of forensic model." Their experiment also included tasks of other kinds. By using open trials and varying the quality of the recording, the number of speakers, the number of clue words, and the lapse of time between known and unknown utterances, the researchers estimated error rates of 6 percent false identifications and 13 percent false eliminations.

In another identification experiment, Tosi studied the effect of the number of speakers on the error rate [6]. With five speakers, the error rate was one percent; with fifty speakers, the error rate increased to 5.7 percent. In the latter trial the operators were forced to decide within fifteen minutes. If the operators were allowed to suspend judgment in doubtful cases, 74 percent of the cases were resolved with 2 percent false identifications and 5 percent false eliminations.

Other researchers showed different results. In one instance, a closed set of five speakers gave 21.6 percent errors for words spoken in isolation and 62.7 percent errors for words taken from context [7]. In this experiment, the operators were Michigan state policemen trained as operators according to the Voiceprint method.

Another experiment was conducted to compare visual and auditory discrimination [8]. In closed trials with eight speakers, 6 percent errors were obtained by listening and 21.6 percent by spectrogram. If the trials were made open by the inclusion of non-catalogued speakers,

6 to 8 percent of the samples were falsely identified by listening and 31 to 47 percent were falsely identified by their spectrograms.

Procedures for speaker identification using digital computers have also been developed. In one study, average spectral patterns for each of ten speakers were measured and stored in a computer [9]. New utterances by one or another of the speakers were analyzed and a pattern recognition type program was used to find the pattern most similar.

Another extensive and recent report [10], describes a method of speaker identification which is semi-automatic; that is, it requires a minimum of operator expertise, and the decisions are made by the machine. The exercises are also forensically suitable in that different utterances are compared by the computer and it is determined whether or not they were spoken by the same speaker. The system can resolve 25 percent of the test cases with no errors. The authors state that in 70 percent of the cases, the probability of a false decision is less than 1 percent. The remarkably good performance of this system is partly due to the nature of the trials, which were closed and used words spoken in isolation.

Broadly speaking, current methods of speaker identification are deficient in two areas: accuracy and objectivity. The accuracy requirement is self-explanatory -- we would like a method which always gives the right answer! The objectivity requirement is intended to make the system or method independent of the operator. This will be done in the following study by reducing the function of the operator to a very simple task. This function can, in principle, be performed automatically for many identification tasks, as will be shown. The

system need not be dependent on "actual use" trails for justification, since the expected reliability of the system will be developed *a priori* on the basis of the statistical distribution of the identification parameters.

In a review and comparison of speaker identification methods published in 1970, Tosi et al. conclude:

> It may be that, when we have learned much more about the sound features that characterize individual speakers, it will be possible to design an instrument that can be a powerful aid to the eye in voice identification, or even one that can operate *automatically* in a completely *objective* manner [11].

Reference 10 cited above suggested areas in which "state-of-the-art" computer-assisted identification methods might be extended. Many of these proposed areas of study are included in this work.

1.2 Statement of the Problem

The purpose of this research, briefly stated, is twofold. First, to determine under what conditions the parameterization of the speech signal provides a method of identification with a given reliability, and, second, to determine how the reliability of the identification is affected by such variables as the number of samples compared, the time between different samples, and other parameters to be defined later.

To complete this task, it is necessary to obtain a sufficient number of speech samples of good quality, to design the analysis procedure which derives the identification parameters, and to model the various kinds of identification tasks on the basis of the identification

parameter distributions. This thesis indicates satisfactory attainment of each of these objectives. A number of suggestions for further work are inherent in this procedure and are presented in Chapter V. Indications of the manner in which the procedure may be automated or specialized for certain tasks is also given.

The data base for this study is approximately eighty minutes of running English speech comprised of twenty-nine readings of the Rainbow Passage, which is reproduced in Appendix A, by twenty-one speakers. A system of computer programs was used to edit this data and to extract an approximation to the vocal tract impulse response magnitude at selected points, by the method of cepstrum deconvolution. It is shown on physiological grounds that these measurements should exhibit greater variability between speakers than the variability of different utterances by the same speaker. This theoretical conclusion is justified empirically by a further system of programs which use a correlation technique to compare the vocal tract impulse response magnitudes at selected points for the same and for different speakers. The set of correlation coefficients so obtained has a decidedly non-normal distribution.

A by-product of this study is a set of highly accurate quantitative measures of the speech spectrum and the speaker fundamental voice frequency. These measurements are presented in Appendix B.

CHAPTER II

DEVELOPMENT

2.1 Physiological Background

The speech signal contains information pertaining to the content of the speech, and gives information that is characteristic of the speaker. If it lacked the former, we would not understand the speaker; lacking the latter, we would not recognize him. Heuristically, this is obvious; a vocoder is intelligible but unrecognizable (except as a vocoder), while one might easily recognize a familiar voice speaking unintelligibly or in a foreign language and not understand it al all. It is shown in this chapter that it is important to separate the two kinds of information in the interest of enhanced reliability of identification, and that it is possible to do so. The manner of separating the identification parameters uses a signal processing technique which has not heretofore been employed in speaker identification.

The difficulties in extracting the speaker identification parameters are inherent in the nature of the speech generator and in the speech process itself. First, we will examine the speech generator at a physiological level -- the vocal mechanism.

The vocal mechanism may be divided into two parts: an acoustic source and an acoustic formant filter. In voiced sounds, the source is a quasiperiodic glottal pulse generated at the larynx. In unvoiced sounds, the source is frictional noise generated by turbulence as the

breath is forced past a constriction in the larynx, pharynx, or oral cavity; or transient signals generated by a stop in the articulatory system. In either case, the acoustic generator is powered by the pressurized air contained in the lungs.

This study is limited to vowel sounds for two reasons: first, since the glottal source is the most efficient and powerful source. The second reason depends upon certain observed features of speech signal and will be developed subsequently.

A schematic diagram of the vocal mechanism is presented in Figure 2.1. It shows the principal parts of the apparatus and outlines the functions which each performs in the speech task.

A prosodic variable is one which depends upon the manner of speech. The most important prosodic variables are intensity (which is related to the acoustic power output) and pitch (which is related to the fundamenal frequency). It is desirable to remove the effect of prosodic variables from the speech signal before an identification is attempted. The intensity and pitch of the speech also show intra-speaker variations due to physiological causes. The system of identification should be isolated from these effects as well. The following paragraphs discuss the ways in which such changes manifest themselves in the measured acoustic signal.

The glottal pressure waveform was indirectly measured by Timcke, von Leden, and Moore, who measured motion pictures of the vocal folds during phonation [1]. They plotted the distance between the vocal folds, as a function of time. Systematic variations were found as a function of pitch and intensity for normal speakers. In the range of pitches

used in normal speech, the higher intensities were associated with pulse-like opening curves and the lower intensities with triangular opening curves. It is reasonable to expect wide variations in the glottal pressure waveform since the shape of the glottal opening curve varies widely. Schematic indications of the vocal tract opening for different conditions of phonation are shown in Figure 2.2.

Stevens [2] writes concerning the glottal pressure waveform itself:

> . . . if the glottal vibration is periodic, the spectrum amplitude of the volume flow is, of course, a line spectrum . . . The shape of each glottal pulse varies somewhat with fundamental frequency and vocal effort for a given talker.

Figure 2.3 shows an approximate "typical" glottal pressure-wave spectrum derived from work by Stevens.

Normal speakers show pitch fluctuations on a cycle-to-cycle basis. Liberman [3] defines the "perturbation factor" as the percent of a large number of cycles of phonation which differ from one cycle to the next by 0.5 millisecond or more. A nonpathologic speaker with an f_0 of 250 Hz would typically have a perturbation factor of 5 percent; one with an f_0 of 100 Hz would normally have a perturbation factor of 25 percent.

The vocal tract acts as the acoustic filter or formant generator in voiced speech. Interest in parameterization of the vocal tract first arose from studies in speech bandwidth compression. If one could encode the speech signal into a small number of slowly varying parameters, one could transmit them on a narrower band than could be done with the original signal. The pitch signal is band limited to about 300 Hz, and

the vocal tract changes during articulation occur on a time scale considerably longer than the pitch period.

Stevens and House have done several studies of vocal tract modeling. In 1955, they reported a three-parameter model that ". . . produces idealized vocal tract configurations which are descriptive of human vowel articulation" [4]. They later designed an electrical analog of the vocal tract to test the performance on their model -- with excellent results.

Steinberg and French employed the term "hub" to describe formant positions [5]. The hub was defined as ". . . the visible or hidden position of bar (formant) 2 of any sound when the sound is made alone." Two facts were recognized: first, that the hub must in some cases be deduced from its effect on adjacent regions, as no visible feature is present; and second, that a large part of the speech signal consists of "transitional patterns" whose formants are displaced by various amounts in different directions, depending upon the nature of the contiguous sounds.

The behavior of the formants is of considerable importance in the detection of consonants. Indeed, it is possible to remove the consonant from a speech segment, leaving only the transition, which will itself cue a listener to hear the missing consonant [6]. S. Ohman, the author of this study notes, however, ". . . that the cost of removing a final segment from a VC utterance or an initial segment from a CV utterance will depend very much on the particular acoustic properties of the sounds that enter into the utterance." (Symbolic notations such as CV are defined in the list of symbols.) The term coarticulation was introduced to name the variation of articulation from values obtained in

neutral contexts or in isolation, which are dependent upon the context in which the articulation is measured.

S. G. Ohman studies VCV form utterances in English speech [7]. He stated that consonant and vowel generation in English involve different uses of different parts of the articulatory mechanism. The VCV utterance is neurally coded as a V-V transition on which is superimposed a command to generate a consonant. He observed that both the stationary portion of the vowels and the transition regions on each side of the consonant were affected by the vowel on the other side of the consonant. This finding is a variance with the older "locus" theory of Libermann et al., in which it was hypothesized that the transition region will tend toward a locus which is characteristic of the interposed consonant. Oh[illegible]et of loci to be associated with each consonant, one for eve[illegible]ir of surrounding vowels.

Stevens and House in 1963 published a study of formant frequency perturbation by consonantal context [8]. They found certain contexts to have a minimal effect on the articulation. These contexts were termed "null" contexts. They conclude:

> . . . the /h--d/ context has a negligible effect on the articulation during the central portion of the vowel. That is, the vowel in the context /h--d/ is generated with essentially the same articulatory configuration as the vowel in isolation.

In an effort to secure maximum isolation for running speech, all vowels employed in this study are isolated on each side by voiceless consonants. An example is the work "path," which is taken from the

context "its path high." Here the desired vowel is isolated on one side by two voiceless consonants and on the other side by three voiceless consonants.

Stevens and House also measured values of f_1 and f_2 in fourteen different contexts. The values of f_1 were little affected by context except in the vowel /ʌ/. This observation, and other anomalies concerning /ʌ/, were claimed by the authors to be due to atypical data from one of the talkers, whose pronunciation of /ʌ/ differed markedly from one context to another. The value of f_2 was nearly systematically affected, with front vowels having f_2 lowered and back vowels having f_2 raised. The authors conclude ". . . the consonantal context has the effect of shifting f_2 from a value appropriate for the null environment toward a more central position."

In summation, fundamental frequency shows perturbations in isolated speech segments due to imperfections in the speech mechanism, or to disorder in the organic or neurologic functions. The glottal pressure wave and the fundamental frequency are also strongly affected by prosodic elements in speech. These changes may occur from one glottal cycle to the next.

Isolated speech tasks, when repeated, show stable articulatory configurations. In context, articulation shows transitional features or coarticulation as well as superposition of different vocal tasks which do not require simultaneous use of the same articulators. Articulatory adjustments are made on a time scale considerably longer than the pitch period.

2.2 Model and Hypothesis in Speaker Identification

It has been shown that the vocal tract configuration is less likely to undergo short term fluctuations than other articulators contributing to the speech signal. Indeed, though the glottal pressure wave varies the driving function, it is the vocal-tract resonances which determine the formants and therefore the vowel which is heard. The following hypothesis is the basis of the proposed method of speaker identification and classification.

First, that contexts in which a vowel is isolated in each side by voiceless consonants protect the vowel from coarticulation by nearby vowels, though allowing coarticulation by the adjacent consonants, and that such contexts are easy to locate by inspection of the speech signal.

Second, that the approximate vocal tract impulse response magnitudes derived from these contexts are sufficiently invariant for a given speaker, and different enough for different speakers, that they serve as speaker identification parameters which may be used to distinguish between speakers in a statistically significant manner.

Since the perceived pressure signal is the convolution of two time functions, the glottal pressure wave and the vocal tract impulse response (and also a "system function" including radiation effects and the recording and analysis process, which for purposes of discussion is ignored), it is difficult to determine directly from the spectrum of the speech signal which parts or features are due to the glottal pressure wave and which to the vocal-tract impulse response. Since the glottal wave is known to vary as a function of the rate of speech, intensity,

and pitch, the difficulty of extracting the vocal-tract impulse response from the speech spectrum, by inspection, is considerable.

Under certain conditions, which are fairly well satisfied in the speech signal, the operation of deconvolution may be performed by cepstrum analysis.

2.3 Computation of the Cepstrum

The cepstrum of a signal is defined as the Fourier transform of the log magnitude spectrum of the signal. Its desirable properties are inherent in the nature of the Fourier transform and in linear-system theory. The following is a partially heuristic development of the cepstrum and its uses.

Let the lower-case letters denote time functions and the uppercase letters the corresponding frequency functions which are Fourier transform pairs:

$$S(f) = F_f[s(t)] \quad \text{and} \quad s(t) = F_t[S(f)] \ .$$

Consider a signal $s(t)$ which is the convolution of two time functions, $g(t)$ and $h(t)$:

$$s(t) = g(t) * h(t) = \int_{-\infty}^{\infty} g(t)\, h(t - \tau)\, d\tau \ .$$

The spectrum of $s(t)$ will be a product:

$$S(f) = F_f[g(t) * h(t)] = G(f)\, H(f)$$

and the log magnitude spectrum of $s(t)$ will be a sum:

$$\log|S(f)| = \log|G(f)| + \log|H(f)| .$$

The inverse transform of this sum of functions is also a sum:

$$F_t(\log|S(f)|) = F_t(\log|G(f)|) + F_t(\log|H(f)|)$$

and defines the cepstrum. The cepstrum, as defined, is a complex function. It is unnecessary to preserve the complex nature of the cepstrum, however, since it is derived by the Fourier transform of a real quantity. The log magnitude spectrum is real by definition, and the symmetry of its Fourier transform, the cepstrum, is therefore known. It is possible, but unnecessary for the purposes of this research, to define a complex cepstrum on the basis of a complex logarithm of the spectrum. This approach is useful where the object is reconstruction of the time waveform, as in echo removal.

The essential nature of the cepstrum is in the Fourier transform of the logarithmic spectrum. This replaces a time domain convolution with a time-like domain summation. The author of this thesis is not aware of any single standard definition of the cepstrum; each writer in signal processing preferring his own. In this thesis, the unqualified term "cepstrum" will refer to the magnitude of the Fourier transform of the log magnitude spectrum of a signal: that is,

$$\text{cepstrum}[s(t)] = |F_t(\log|F_f[s(t)]|)| .$$

Where it is not intended that the magnitude of the cepstrum be taken, the term "complex cepstrum" will be used. The complex cepstrum derived from the complex logarithm, though not used in the thesis, will be explicitly described where it is mentioned.

Suppose that $s(t)$ is periodic. Then $S(f)$ will be a line spectrum and will have peaks spaced at equal intervals in frequency. The function $S(f)$ is therefore periodic in frequency, and its cepstrum will be a line cepstrum.

The cepstrum of a voiced speech segment will be the sum of a continuous function of time represented by $F_t[\log|H(f)|]$, since the vocal tract frequency response $H(f)$ varies slowly and aperiodically with frequency, and a line cepstrum represented by $F_t[\log|G(f)|]$, since the glottal pressure wave spectrum $G(f)$ varies rapidly and periodically with frequency. Therefore, under the above-stated restrictions on $h(t)$ and $g(t)$, their respective contributions to the cepstrum of $s(t)$ may be readily distinguished. The sharp peaks representing the voice harmonics in the linear spectrum are smoothed by the log magnitude operation to an approximately sinusoidal form. The contribution of $g(t)$ to the cepstrum will therefore be seen principally as a few peaks spaced at intervals of the pitch period in the cepstrum domain.

Figure 2.4 presents the time waveform of a 50-millisecond segment of voiced speech, comprising 512 samples. The pitch period is indicated by τ, and is the interval between successive glottal pressure-wave cycles.

The Fourier transform or linear spectrum magnitude of the speech segment is shown in Figure 2.5a and the log magnitude spectrum is shown in Figure 2.5b. In each case, only 256 points are shown and they correspond to the positive frequencies. It is apparent that the log magnitude operator has smoothed the peaks which represent the voice harmonics to an approximately sinusoidal form. The voice harmonics are positioned at intervals of $f_0 = 1/\tau$ in the frequency domain since the glottal pressure wave is quasi-periodic. The particular form of the discrete Fourier transform which was used in this analysis gives frequency points at 10, 30, 50, . . ., 5110 Hz.

The cepstrum, or inverse transform of the log magnitude spectrum, is shown in Figure 2.6. The periodicity of the log magnitude spectrum is telescoped in the cepstrum into the pitch peak. The total length of the cepstrum corresponds to a pitch period of 25 msec. Since the cepstrum is the inverse transform of a frequency function, its domain is a time-like domain. The cepstrum is related to the autocorrelation function, which is the inverse transform of the square of the spectrum, in that it tends to emphasize periodicity in the spectrum of a signal. The original pitch period of the speech signal under discussion was τ. The primary pitch peak occurs at $\tau = 10$ msec in the cepstrum. It is not unusual for a smaller pitch peak to be seen at $2\tau = 20$ msec, as is shown in Figure 2.6. The successive pitch peaks in the cepstrum have been termed "rahmonics" of the signal.

The position of the pitch peak may therefore be related to the fundamental frequency of the speaker. For the particular analysis used in this study, the fundamental frequency corresponding to a pitch peak

in cell N is 10240/N Hz. Choosing N = 41 gives an f_0 of 250 Hz, which is above any fundamental frequency likely to be used by an adult speaker. Since the pitched components of the speech signal are telescoped into the pitch peak, and its rahmonics into the cepstrum, the portion of the cepstrum from the origin to the 41st point may be considered as an approximation to the Fourier transform of the vocal tract frequency response. This is not strictly accurate. The low "quefrency" part of the cepstrum, as the cepstrum independent variable is called, is the Fourier transform of the vocal tract frequency response multiplied by the glottal pressure wave; therefore, the term "approximate vocal tract impulse response" is used to name the low-quefrency part of the cepstrum. If the pitch peak is suppressed and the remainder of the cepstrum is Fourier transformed, a curve similar to Figure 2.7 is obtained. This curve is the vocal tract frequency response multiplied by the glottal pressure wave spectrum. The use of the cepstrum is to remove the rapid oscillations in the spectrum which are due to the quasi-periodic nature of the driving function.

Appendix B contains mean cepstral pitch determinations and mean vowel spectra which were derived by the technique outlined above.

2.4 Derivation of the Performance Prediction

The identification or classification task may be modeled as a random process that may respond to an input condition a or an input c , wherein it is desired to determine whether the input is a or c from an examination of the output. The decision-making process is shown schematically in Figure 2.8.

It is assumed that the mean output in the presence of condition a is greater than that in the presence of condition c , and that the distributions of responses to the possible a and c inputs are known. In essence, a threshold T is chosen, and it is assumed that if the output is greater than T , the input is a , and if the output is less than T , the input is c .

In the discussions to follow, condition c will be associated with the cross-correlations, i.e., with the absence of the desired speaker characteristic, and condition a will be associated with the autocorrelations, in which the desired speaker characteristic is present. Though reference is made to speaker characteristics, the approach is perfectly suited to identification as well as classification and will be so applied -- for surely the identity of a speaker is one of his most important characteristics.

A threshold T may be selected that gives any desired probability of detecting either of the conditions at the expense of the likelihood of detecting the other, for a pair of distributions such as are shown in Figure 5.1. For example, if the threshold were set at the mode of the c distribution, the probability of detecting a would be quite high; but the probability of detecting c would be only 50 percent, or chance level.

Area (a) is termed the false identification region; that is, the area in which a cross-correlation coefficient is above the threshold. Area (c) is termed the false elimination region, in which an auto-correlation coefficient is below the threshold.

Let the probability of a true decision be p_t and that of a false decision be p_f. Values of p_t and p_f may be derived from the actual coefficient distributions. Assuming that the coefficient distribution obtained by the use of the programs SPKTST or GRPTST is a fair representation of the universe and that selection from the universe occurs randomly, then the probability of obtaining a coefficient between certain limits is the ratio of the area under the coefficient distribution between those limits, to the area under the entire curve. This function is tabulated for normal distributions; however, it will be shown that the means and standard deviations derived from the correlation coefficient distributions in this thesis do not adequately represent the actual distributions.

A reasonable choice for the threshold location is to place it between the distribution means in such a way that the probability of false identification equals the probability of false elimination. This choice allows a single value of p_t to give the probability of correct identification or elimination, and a single value of p_f to give the probability of false elimination or identification. Under the above definition, $p_t + p_f = 1$, always.

It is assumed that the identification or classification task consists of n independent experiments, each of which is the comparison of a unique pair of utterances. The theoretical probability of success in one experiment is p_t, where success is defined as the correct identification or elimination of a sample, and the theoretical probability of failure in one experiment is therefore p_f. The criterion for a decision on the basis of n experiments should be set on the basis of the number of experimental results greater than T (which indicate

identity of speakers for the two utterances) and the number of results less than T (which indicate difference of speakers for the two utterances). The most stringent criterion is to require n results greater than T for identification and n results less than T for elimination. In general, the predicted results will be the terms of the binomial expansion,

$$P(m,r) = (p_t)^m (1 - p_t)^r = (p_t)^m (p_f)^r ,$$

wherein $P(m,r)$ is the probability of obtaining m true responses and r false responses in $m + r$ experiments, where $m + r = n$. Only the end terms, $P(n,0)$ and $P(0,n)$, will be considered determinate results. Any task in which some experiments have results greater than T (identify) and some have results less than T (eliminate) will be considered indeterminate. The predicted probabilities for different kinds of decisions based on n experiments may accordingly be written:

$$\text{correct identification or elimination} = (p_t)^n = P_t(n) ,$$

$$\text{false identification or elimination} = (p_f)^n = P_f(n) ,$$

and the

$$\text{indeterminate result} = 1 - (p_t)^n - (p_f)^n = P_i(n) ,$$

where $P_t(n)$, $P_f(n)$, and $P_i(n)$ denote the probabilities of true, false, and indeterminate results, respectively, of an identification task consisting of n trials.

While it would be desirable to design a system that resolves nearly all of the identification or classification tasks correctly, it

is important to determine what proportion of the determinate tasks are correct. The quantity $R(n)$, defined by,

$$R(n) = p_t^n/(p_t^n + p_f^n) ,$$

indicates the reliability with which determinations are made by this method. The value of $R(n)$ gives the ratio of correct determinations to all determinations as a function of p_t, p_f, and the number of experiments, n.

CHAPTER III

EXPERIMENTAL PROCEDURE

3.1 Data Collection

The primary data used in this thesis consists of magnetic tape recordings of different speakers reading the Rainbow Passage. Three one-week recording sessions were held in 1973, 1974, and 1975.

The exact equipment configuration which was used in the 1973 recording session was not under the author's control and cannot now be determined; however, the following three items include all equipment which was employed: tape recorder--audio, Nagra Model II S/N PH06710602, used in 1973, 1974, and 1975; sound isolation booth--ISA Model 40 S/N 328, used in 1973 and 1974, and a field-modified telephone booth, used in 1975 (a report on the properties of the booth is included in this thesis as Appendix C); and microphone--Electrovoice Model 664, used in 1973, Sennheiser Model MD421U, used in 1974 and 1975.

Over 200 recordings were obtained, including multiple readings by the same speaker. All of the data were transferred from individual reels of 1/4-inch audio tape to a 1/2-inch, seven-track instrumentation recorder format. The Nagra recorder was used for playback and an Ampex Model FR 1300, S/N 6480126, was used for re-recording. Channel 1 of

the instrumentation tape was devoted to a master timing signal which allows the desired speech signal to be accurately located. This tape is referred to as the Voice Analysis Master tape (VAM).

The data group used in this thesis is a subset of the recordings. Nine male and twelve female speakers were chosen who had made a total of twenty-nine readings. The pertinent data concerning these subjects was obtained by questionnaire at the time of the recording (see Table III.1). The selected readings were transferred from the VAM to digital magnetic tape in a format suitable for input to a digital computer. The timing signal was used to control an identification channel on the digital tape so that the digital and analog tapes could be synchronized.

The analog-to-digital recording procedure is diagrammed in Figure 3.1. The VAM is positioned to the desired sample reading. The speech signal is low-pass filtered (5.0 kHz) and an analog-to-digital converter (operating at 10.24 kHz) generates a digital representation of the filtered speech signal. The timing marks on the VAM occur at one-second intervals and are used to advance the identification counter, whose value is recorded on the digital tape simultaneously with the data.

The upper-frequency cut-off of 5.0 kHz was chosen to limit the bandwidth of the data to a reasonable representation of the speech signal. The sample rate of 10.24 kHz was chosen so that 512 real samples represent a time interval of 50 msec, which places the frequency domain samples at intervals of 20 Hz for computational convenience.

Twenty-nine readings were transferred to digital tape in this manner. The repeated readings were made at time intervals ranging from

a few minutes to two years. This data was used to determine the stability of the identification parameters.

3.2 Generation of Speech Cepstra

The second step in the identification process is the selection of certain speech segments and the computation of their cepstra. The selection in this instance was made by manual comparison of transcripts and computer output, but there is no theoretical reason that the selection cannot be controlled by a computer with the appropriate program and peripheral equipment.

An analog recorder was used to convert each speech signal to a chart recording. Figure 3.2 is a chart recording of speaker 111 reading a part of the "Rainbow Passage." When the recording was made, index marks were made on the lower edge of the chart to show breaks in the phrasing; those on the upper edge mark the speech segments of interest. In this thesis, only vowels separated on each side by voiceless consonants were studied. The spaces isolating "take" and "shape" can be seen.

Once the desired speech segment is located, the cepstra are obtained in the following way. A block of 512 real samples, which is considered to be the first recognizable glottal cycle, is chosen at the beginning of the segment. This block is Hanning weighted, the window function being defined by

$$W(n) = 0.54 - 0.46 \cos(2\pi n/512) \quad .$$

The Hanning window was chosen to minimize the side-lobe level. The Fourier transform of the windowed time function contains 256 complex

frequency points. The log magnitude operation gives the spectrum at 256 real frequencies, and this spectrum is inverse Fourier transformed to yield the complex cepstrum. It is possible to define the cepstrum on the basis of complex logarithms of the spectrum, but this is not considered necessary for this work. The magnitude of the cepstrum is retained for use in identification and simulation tasks, and the complex cepstrum is retained for use in deriving voice pitch and spectral information.

Successive Hanning-weighted blocks of a signal are statistically independent if they are separated by one-half the block length. Accordingly, successive blocks of 512 real samples, each 50 msec in length, are taken at intervals of 25 msec after the first sample until the entire segment has been processed. Thus, the number of cepstra derived from a single utterance will depend upon the length of the utterance. Cepstra from within the same utterance by a given speaker have not been found to be significantly more similar than cepstra from different utterances by the same speaker. Consequently, the average cepstrum for each utterance is considered by the subsequent analysis to be representative of the utterance.

The first computer program (EDITVA) allows the operator to examine data from the digital tape and to obtain a permanent record of the waveform at any point. This program was used to check the quality of the digital recording and to locate the speech segments of interest for further processing.

Once the desired speech segments were located, a work sheet (Figure 3.3) was prepared. This sheet gives the digital ID number for

locating each segment to be analyzed. Once the working sheet has been prepared, the second analysis program (EXTRAC) may be started.

The second program allows data to be selected from the digital tape, and the cepstra of the selected segments to be recorded on another digital tape, called the "cepstrum tape."

The EXTRAC program generates two forms of output. The first output form is shown in Figure 3.4. The program requests the operator to enter the time (in milliseconds) within a given ID number at which each speech segment of interest begins. These requests appear along the left edge of the display. For the second output form (Figure 3.5), the program computes and displays the log magnitude spectrum and the cepstrum of each of the requested segments. After each cepstrum is computed and displayed, the operator may choose either to have it ignored by the program, or to have it labeled and included on the cepstrum tape.

The cepstrum tape generated by EXTRAC contains cepstra and their associated indexing information: the subject number, vowel number, the digital data tape ID number, the start time within the ID, and an index number which allows the actual program listing to be determined. Each time the second analysis program (EXTRAC) is started, the index number is reset to 1, and each operation is assigned a unique index number within the run of the program. A part of the reference listing of the cepstrum tape is shown in Figure 3.6.

3.3 Classification and Identification Programs

The term auto-correlation refers to a correlation coefficient between cepstra which share any common characteristic. For example, in an identification experiment, coefficients derived from different

utterances by the same speaker would be considered auto-correlation coefficients. In a classification by sex experiment, all coefficients derived from subjects of the same sex would be considered auto-correlation coefficients. The term cross-correlation refers to correlation coefficients between cepstra which do not share the common characteristic of interest.

The correlation coefficients are derived in the following way. The user first selects as many groups of cepstra as are desired, each of which may contain one or more cepstra. The mean value of each group is reduced to unity. All groups are summed and the sum is reduced to unity mean. This produces the average cepstrum for the experiment. The average cepstrum is subtracted from each of the individual group cepstra, thus converting the group cepstra to the deviations of each group cepstrum from the mean cepstrum. Finally, the correlation coefficient for each pair of group cepstrum deviations is computed. If the two group cepstrum deviations share the common characteristic of interest, the resulting correlation coefficient is regarded by the program as an auto-correlation coefficient; if they do not share the common characteristic, the coefficient is regarded as a cross-correlation coefficient. The group cepstrum deviations from the mean cepstrum are correlated, rather than the group cepstra themselves, in order to emphasize most strongly the differences between the cepstra. The group cepstra themselves correlate with one another with coefficients almost always greater than 0.95, whether or not they were uttered by the same speaker or by members of the same group.

Various kinds of classification and identification tasks may be simulated by the computer programs SPKTST and GRPTST. Each of these programs accepts the cepstrum tape generated by EXTRAC as input and requests instructions from the operator via a graphics display terminal concerning the operations to be performed. The programs perform identical processing and generate output of the same form. They differ in the degree of flexibility given the operator in choosing the nature of the experiment.

Program SPKTST assumes that cepstra are to be distinguished on the basis of subject number. This allows speaker identification tasks to be simulated. This program also generates two forms of output: the first (Figure 3.7) gives a condensed listing of the contents of each cepstrum group; the second (Figure 3.8) lists the statistical parameters derived from the distribution of the correlation coefficients, and a histogram of their distribution. The operator may request more detailed information for any of the output forms, including a listing of all the correlation coefficients as they are generated. Cumulative statistical information is also stored by the program and is displayed at the end of the run or upon operator request.

Program GRPTST allows cepstra to be grouped for study in an arbitrary manner. This program is used to obtain the coefficient distributions for speaker classification experiments. The output of GRPTST is identical to that of SPKTST, and the input protocol differs only in that the program allows the user to specify a code to be used in classifying each cepstrum as an auto-correlation or cross-correlation.

CHAPTER IV

RESULTS

4.1 Speaker Identification

Program SPKTST was used to obtain the distribution of correlation coefficients and associated statistics for use in speaker identification simulation. Each of the four classes of speakers, which are male smokers, male nonsmokers, female smokers, and female nonsmokers, had seven vowels processed.

For each vowel within each class, the sum of all cepstra derived from each utterance and reduced to unity mean was used as the representative cepstrum for that utterance. Therefore, the auto-correlation coefficients give correlations between different utterances of the same vowel by the same speaker, and the cross-correlation coefficients give correlations between utterances of the same vowel by different speakers. For each class of speaker there is one pair of correlation coefficient distributions and associated statistics for each of the vowels.

Table IV.1 shows summary information about the identification experiments, and lists for each vowel within each class the number of samples, the distribution mean, and the standard deviation for each of the coefficient distributions.

The data presented in Table IV.1 show that certain vowels have μ_a , the auto correlation mean, significantly greater than μ_c , the cross correlation mean for each of the four speaker groups, and are in

this sense good identification parameters. Other have μ_a and μ_c too close together or fluctuate from one group to the next and are relatively poor identification parameters. It is reasonable to expect that certain of the vowels should fail consistently. Vowels 2 and 12 are often embedded within an utterance where they are very difficult to locate accurately. It is also probable that many cepstra identified by either of these vowel numbers are labeled in error. Vowel 102 is a diphthong, occurring in the words "type" and "sky." Unlike the other diphthong used in this study (number 32 taken from "take", "shape", and "cate"), vowel 102 changes its spectrum radically during the course of the vowel. It is a very obvious diphthong, and samples from one part of an utterance correlate poorly with other utterances. For these reasons, only four of the vowels, 4, 5, 8, and 32, are considered acceptable for identification purposes. They are discussed in the following paragraphs.

Table IV.2 lists the statistical parameters of the summed distributions for each of the speaker classes. Each of the distributions includes all four of the "good" vowels. A X-square test was employed to test the likelihood that the observed coefficient distributions are drawn from a normally distributed universe. In general, the fit is very poor; therefore, the use of the standard deviation is not strictly defensible.

For each of the speaker groups, and separately for the auto-correlation and the cross correlation distributions, Table IV.2 shows the number of samples, the distribution mean, the standard deviation, and the X-square probability of normalcy of the sample universe.

The data presented in Table IV.2 are derived in the following way. Each vowel or speaker is processed separately, and the distributions and statistical parameters are summed to give the accumulated values. In the test for vowel 4, for example, each of the groups, 1 through 4, was run separately; the resulting distributions were summed to give the distributions shown in Figure 4.1, and the values of Σx and Σx^2 for each group were added together to give the totals used in deriving the accumulated mean and standard deviation. Figure 4.1 shows the auto-correlation and cross-correlation coefficient distributions for vowel 4. To study identifiability by group, each of the four vowels was run separately, and the resulting distributions were summed to give the overall performance. Figure 4.2 shows the auto-correlation and cross-correlation coefficients for Group 1. Figures 4.1 and 4.2 are typical of vowel and group identifications, respectively.

4.2 Speaker Classification

Program GRPTST was used to obtain correlation coefficient distributions and associated statistics for use in the simulation of speaker classification tasks. This program allows very flexible grouping of the cepstra into various classes. The computations performed by the program are the same as those performed by SPKTST. The set of speakers used in this analysis was chosen so that the effects of smoking, sex, and time lapse between readings could be individually studied. The results obtained from the operation of program GRPTST are presented in this chapter. In Chapter V these results will be interpreted and used to predict the reliability and accuracy of the proposed method of classification.

Several modes of classification were studied. First, the program was set to include all samples of a given vowel by male speakers in one group, and all samples by female speakers in another group. The auto-correlation coefficients thus produced represent the correlation between speakers of the same sex (but not necessarily the same speaker). The cross-correlations represent samples of speakers of different sexes. A histogram showing the resulting coefficient distributions is shown in Figure 4.3. This histogram is typical of those obtained in classification experiments.

Next, the program included smokers of one sex in one group and nonsmokers of the same sex in another. In these tests, the auto-correlations are between two smokers or two nonsmokers, and the cross-correlations are between one smoker and one nonsmoker.

Data Group 5 included six readings of the "Rainbow Passage" by speakers who had previously recorded the passage one year or more prior to that included in Group 5. The program was directed to include contemporaneous readings in one group and widely-separated readings in another. Thus, the auto-correlations in this test are between samples recorded by the same speaker at essentially the same time, and the cross-correlations are between samples recorded by the same speaker at very different times.

In all of the above classification tests, the histograms for the four vowels 4, 5, 6, and 32 were determined separately. They were then added and their statistical parameters were grouped to form the distributions.

The X-square probability of sampling from a normal universe was computed for each of the preceding test cases. The resulting P values,

along with the number of samples, the mean, and the standard deviation, for both the auto-correlation and cross-correlation distributions, are summarized in Table IV.3.

4.3 Normal Versus Actual Coefficient Distributions

There is slight justification for applying normal statistics to the speaker identification and speaker classification problems. Out of 24 tests on which χ-square was evaluated, 13 have $P < .001$, three have a "poor" fit with $P < 0.05$, and eight have a "good" fit with $P \geq 0.05$, sometimes much greater.

Table IV.4 gives the number of samples and the type of correlation with the corresponding P values. The auto-correlation distributions seem to be more nearly normal than the cross-correlation distributions. Better results in this sense are also associated with smaller numbers of samples.

An attempt was made to match each side of the coefficient distributions with half of a normal distribution. Values of χ-square were computed for these distributions; however, no better fit was obtained than for the single symmetrical normal distribution. For this reason, the actual coefficient distributions were integrated to give empirical probability density functions. These functions were then used in place of the normal approximations to obtain the probabilities of true and false decisions in identification and classification tasks.

CHAPTER V

CONCLUSIONS

5.1 Speaker Identification

The correlation coefficient distributions obtained by use of the program SPKTST were used to derive values for p_t and p_f in simulated identification tasks. The values of p_t and p_f for each vowel were obtained from the sums of the correlation coefficient distributions of each of the four speaker groups for the given vowel, and the values of p_t and p_f for each group were obtained from the sums of the correlation coefficient distributions of each of the four vowels for the given speaker group. These are shown in Table V.1. For $n = 1$, 2, and 3, and based on the given values of p_t and p_f, values for $P_t(n)$, $P_i(n)$, and $P_f(n)$ are shown in Table V.2, and values for $R(n)$ are shown in Table V.3.

The identification by vowel is seen to be more reliable than the identification by group. This is not unreasonable, as different utterances of the same vowel might be expected to be more similar than utterances of different vowels, whether the utterances were by the same or by different speakers.

For tests involving three utterances, in identification within a group over all vowels, 29 to 50 percent of the test cases are determinate and correct; they comprise 88 to 98 percent of all determinations. For tests involving three utterances for a given vowel, including samples

from all groups, 36 to 46 percent of the test cases are determinate and correct, these comprise 94 to 98 percent of all determinations.

It is seen that, on running English speech, the system operates to give identifications with an accuracy comparable to that obtained by other methods, some of which use isolated clue words rather than running text, whether the identification was based on a visual, auditory, or computer-matched method.

5.2 Speaker Classification

The correlation coefficient distributions obtained by the use of program GRPTST were used to derive values for p_t and p_f in simulated classification tasks. Values of p_t and p_f, derived from the distributions, are shown in Table V.4. Values for $P_t(n)$, $P_i(n)$, and $P_f(n)$ are shown in Table V.5, and $R(n)$ in Table V.6, for $n = 1, 2,$ and 3.

The classification tasks attempted were the determination of sex, distinguishing between smokers and nonsmokers, and distinguishing between contemporaneous recordings and those separated by more than one year.

There is little doubt of the significance of the difference between the means of the classification idstributions. The estimated standard deviations of the means are orders of magnitude less than their differences. The system does not, however, allow an acceptable probability of a correct decision while suppressing false decisions.

For the sex determination, and distinguishing between smokers and nonsmokers, 16 to 25 percent of the tests are determinate and correct,

which comprise 64 to 84 percent of the determinations. The distinction between female smokers and nonsmokers is particularly unclear.

The distinction between contemporaneous recordings and those separated by more than one year was performed separately for each vowel, and the resulting distributions were summed to give the distributions that were used in measuring the probability of true and false decisions. There is a significant difference between utterances of the same vowel at different times by the same speaker, but it is much smaller than the difference between utterances of the same vowel by different speakers. Twenty four percent of the determinations of time lapse between recordings, comprising 81 percent of the total determinations, are correct.

5.3 An Entirely Automatic Identification System

It has been shown that the present system is capable of identifying speakers, but not capable of classifying them. One design for an automatic speaker identification system would be a digital computer with a conversational input/output device such as a teletype, and an analog-to-digital converter with a bandwidth of 5 kHz connected to a microphone. It is assumed that the computer would be programmed with a data base of known speakers which includes all legitimate candidates for identification. An identification exercise would then be conducted in the following manner.

The computer would design a test sentence containing three vowels suitable for identification by a random selection of words in an arbitrary order. This would serve to protect the system from deception by a previously prepared recording. A short list of words suitable for the

sentence framework "verb the adjective noun" could include all four suitable vowels in the verb, adjective, and noun lists. Different sentence frameworks could be stored and one chosen randomly for each test for further deception-proofing.

The computer would then display the test utterance and request the subject to read it aloud. It is in the interest of the subject to read it carefully, to prevent identification failure. The perceived signal would be recorded in the computer's memory and parsed into the speech segments of interest. It is relatively easy to distinguish voiced from voiceless speech segments, and the computer begins with the knowledge of what to look for in the sentence.

Once the vowels have been located, the computer would extract cepstra from each utterance and compare them with those stored in its memory. Should all three samples match the cepstra of a known speaker, a positive identification would be confirmed. Should there be fewer than three matches or a match with more than one speaker, the identification would be denied.

5.4 Suggestions for Further Research

It is possible to work in two directions within the framework of the present system design. On the one hand, it is likely that the system performance can be enhanced by modifications in the computations which are performed and, on the other hand, the available data base is much greater than that used in this thesis, and additional results or different kinds of results might be derived from a larger sample.

One area to examine would be the effect of prewhitening the speech signal before cepstrum extraction. Since the cepstrum is large

for only a small range of values, this operation could also be effectively performed by multiplying the test cepstra by some weighting function before obtaining the correlation coefficients. The present correlation programs, SPKTST and GRPTST, are capable of performing this operation on the data. The correlation programs are also capable of handling any number of inverse frequency points in the cepstrum. The present cutoff of 41 points was chosen so as to eliminate all pitch peaks from the cepstra before correlation, as 41 points in the cepstrum domain corresponds to an f_0 of 250 Hz.

In the data base from which the samples for this thesis were taken, there are recordings which were made by individuals who were attempting to disguise or change their voices. It would be interesting to see if their identifiability is affected by such attempts at voice modification. The data-base logs include the age, state of vocal training, and place of origin of each speaker. Any combination of these could be studied for classifiability, or at least for a significant observable difference.

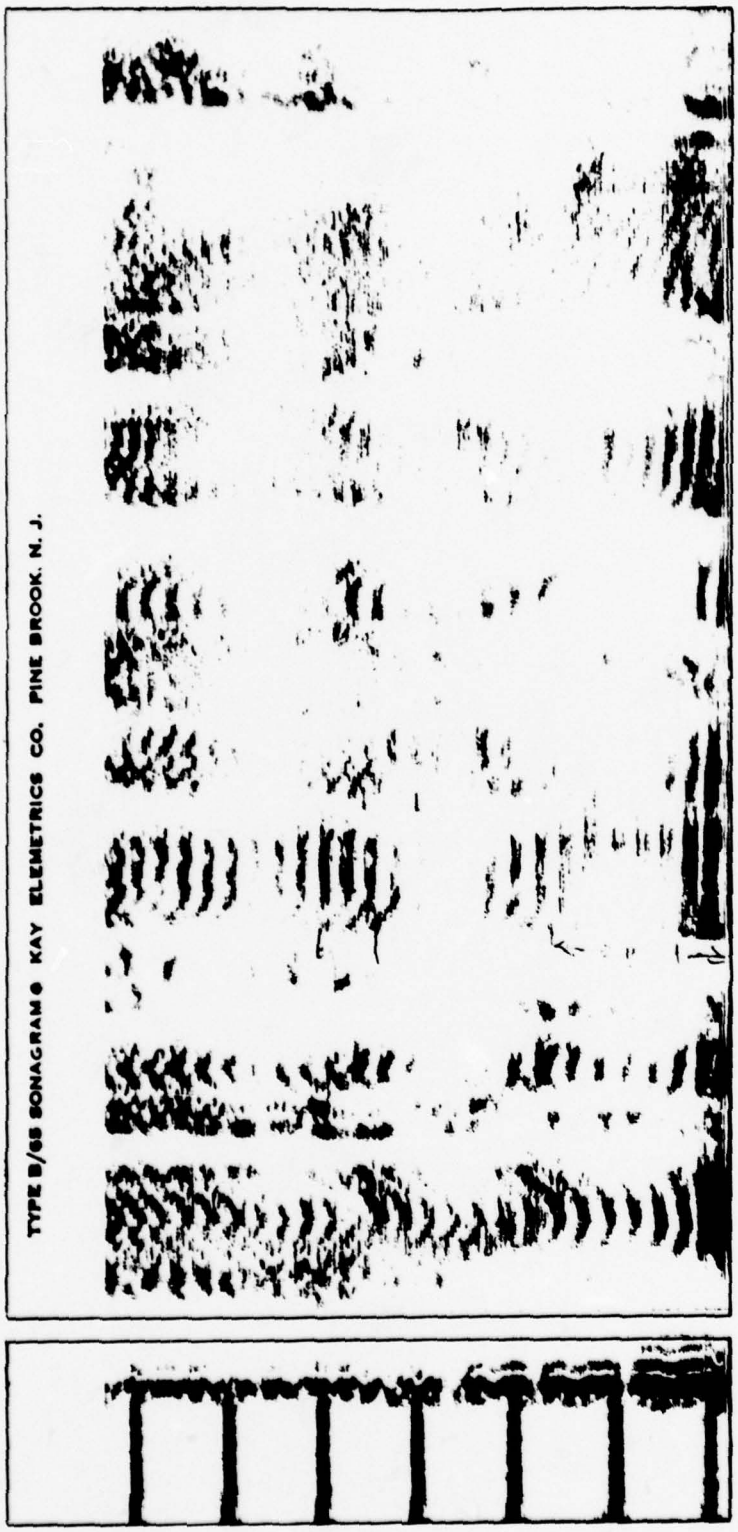

Figure 1.1. Voice Spectrogram.

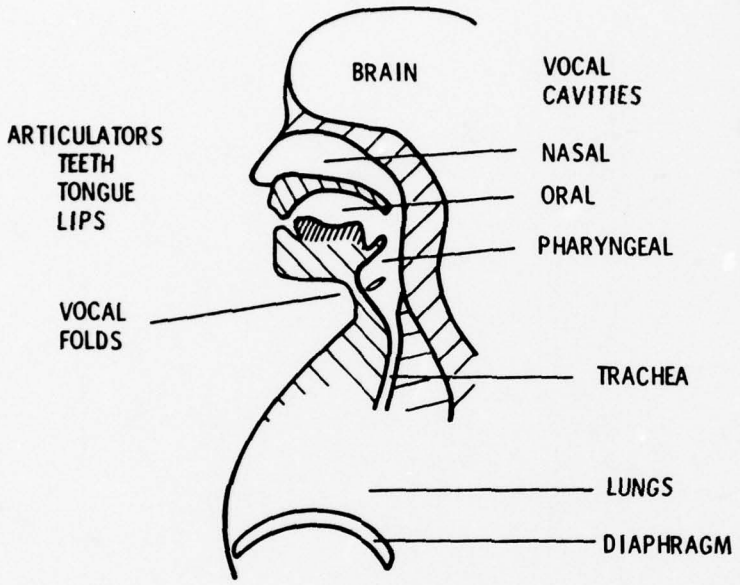

Figure 2.1. Schematic Diagram of the Vocal Mechanism.

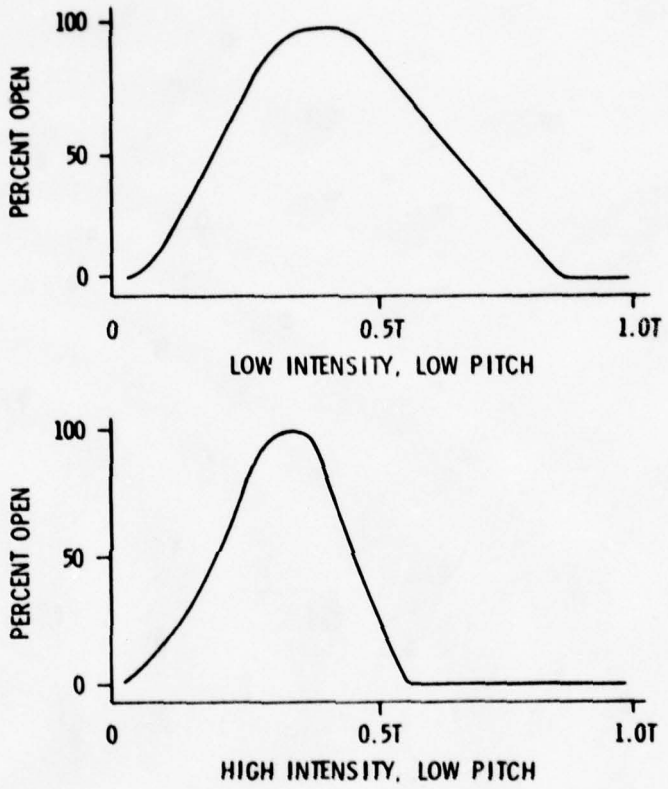

Figure 2.2. Vocal Fold Opening Curves.

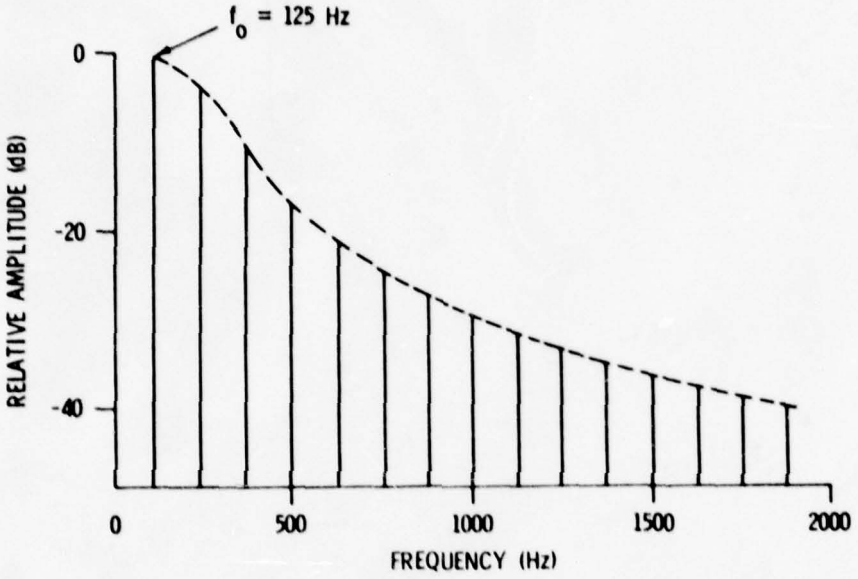

Figure 2.3. Typical Glottal Pressure Wave Spectrum.

Figure 2.4. Time Waveform of 50-Millisecond Speech Segment.

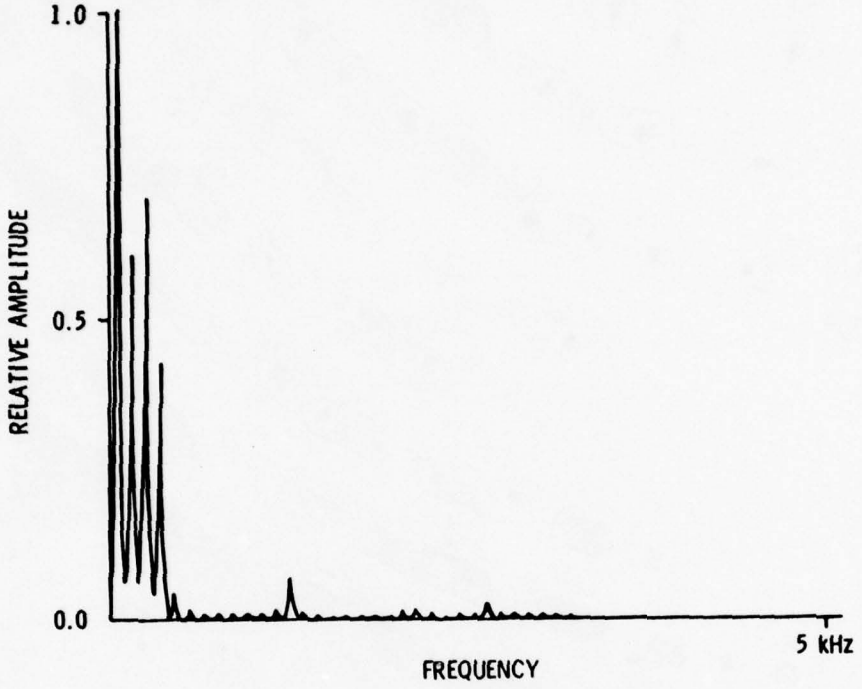

Figure 2.5a. Linear Spectrum.

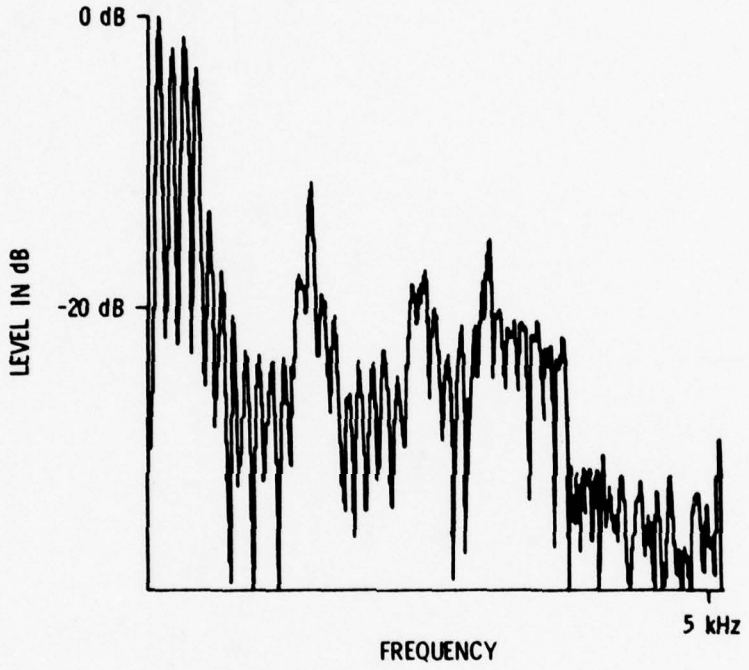

Figure 2.5b. Log Magnitude Spectrum.

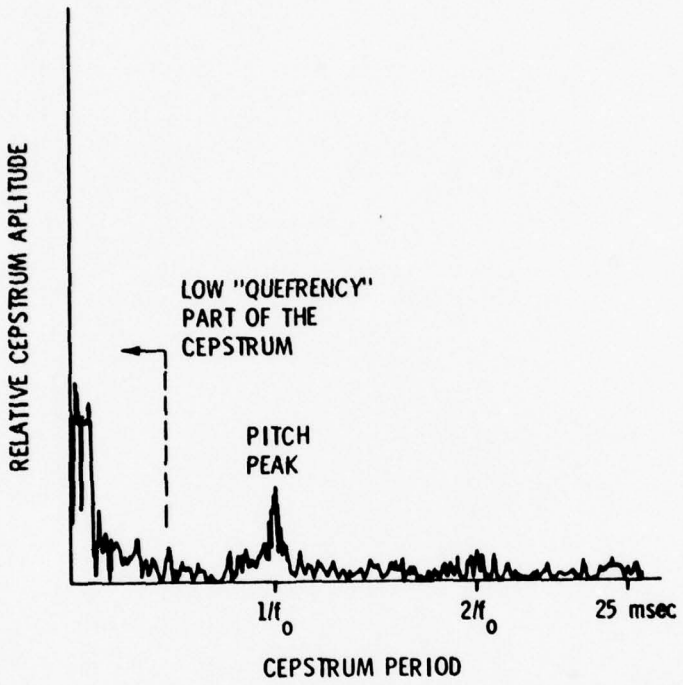

Figure 2.6. Cepstrum.

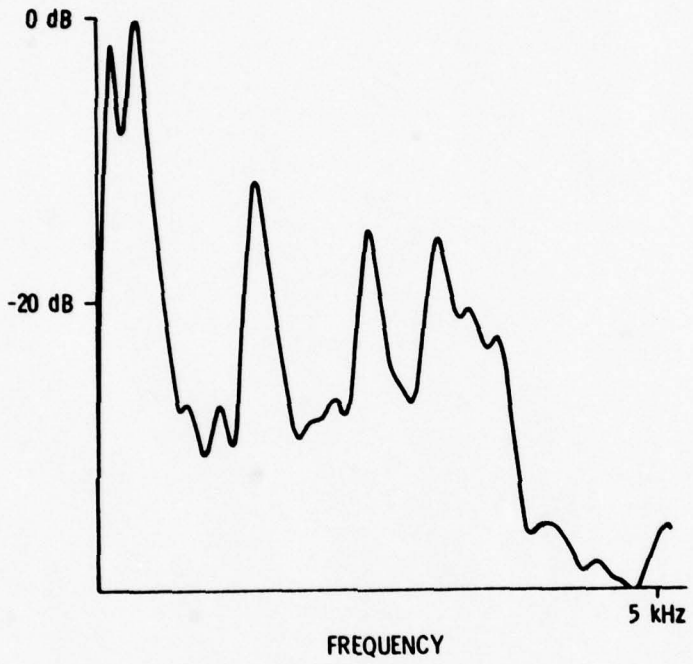

Figure 2.7. Approximate Vocal Tract Frequency Response.

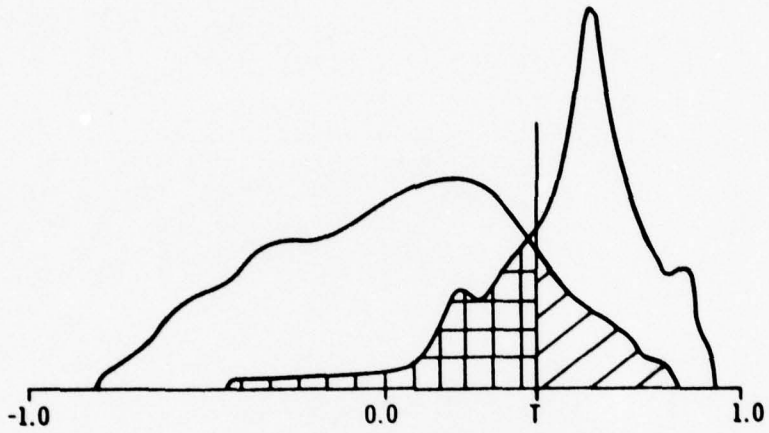

Figure 2.8. Decision-Making System Schematic.

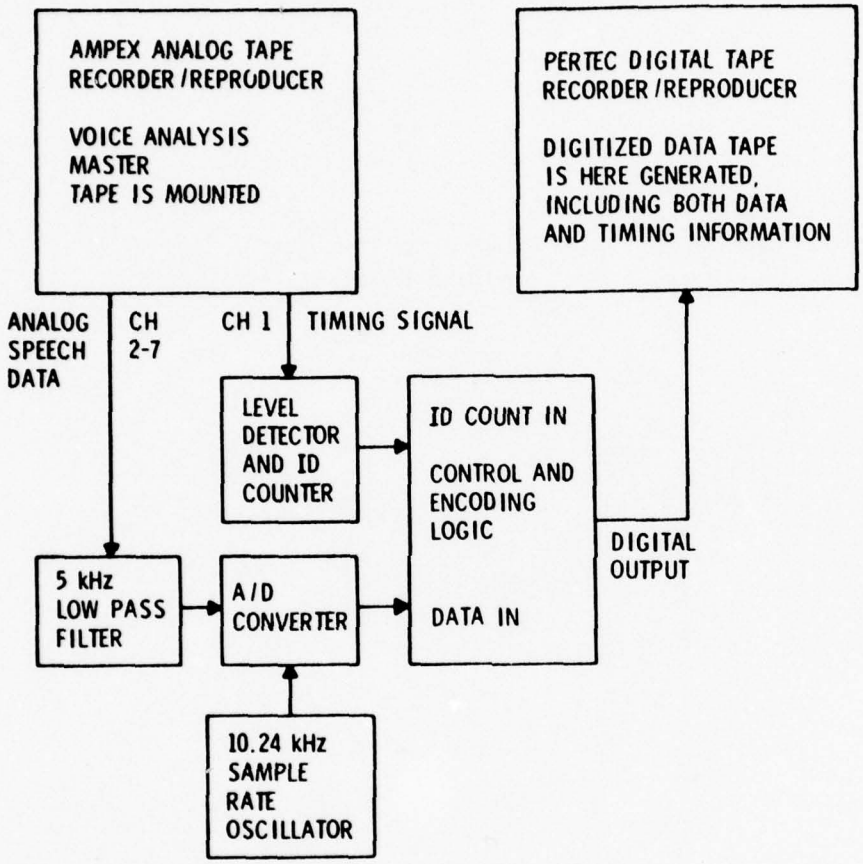

Figure 3.1. Analog-to-Digital Tape Transfer.

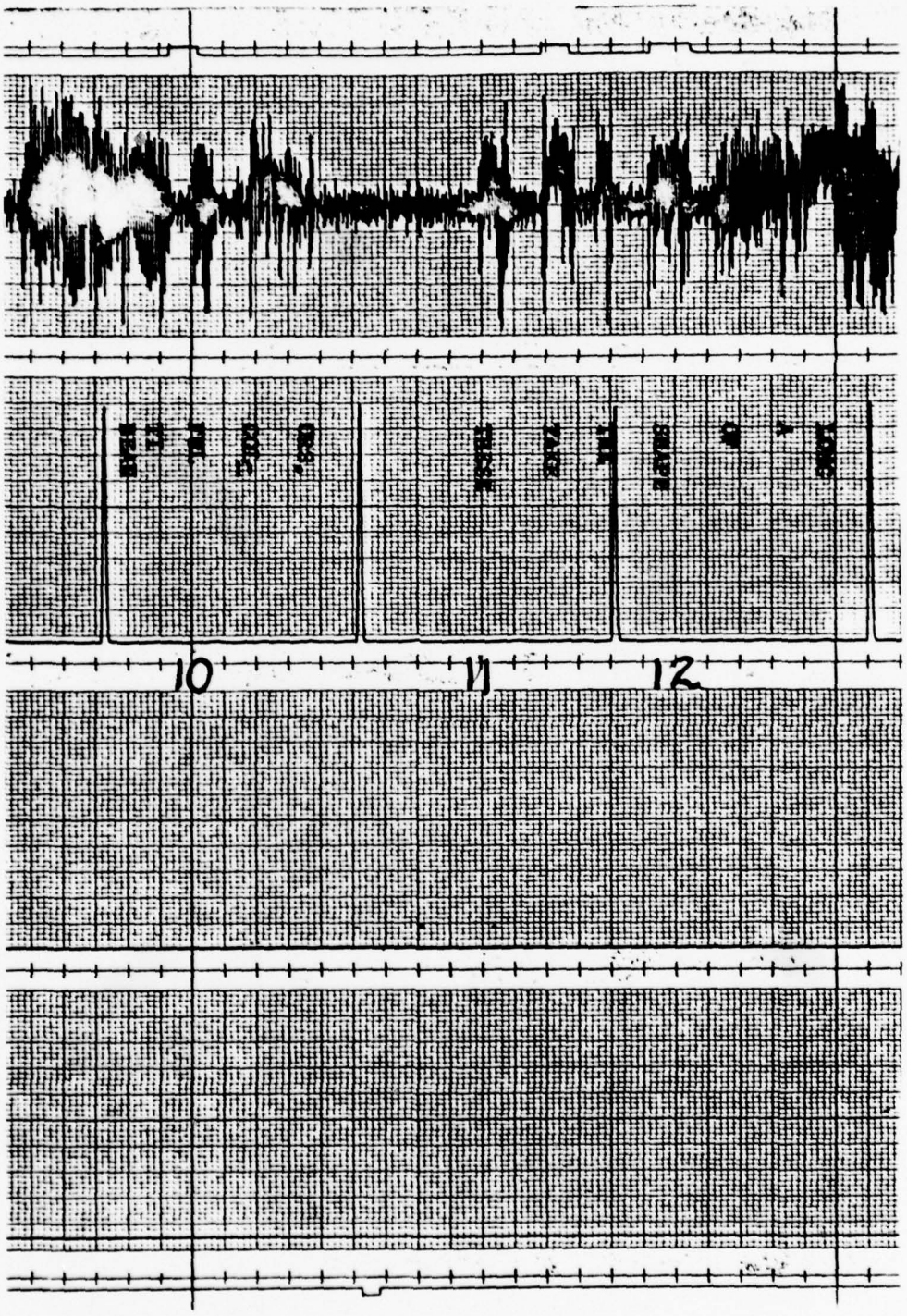

Figure 3.2. Specimen Chart Recording of the Speech Signal.

SPEAKER_____ JOB ID_____ TAPE_____ DATE_____ TIME

NOTES:

LINE	VOWEL	NUMBER	ID	ENVIR-ONMENT	NOTES
3	I' (I)	2		tɪf	
4	EI (eI)	32		teIk	
4	EI (eI)	32		leIp	
5	A (æ)	5		pæs	
8	U" (ʌ)	12		/ʌk	
9	A" (ɔ)	8		pɔt	
11	U" (ʌ)	12		bʌt	
13	EI (eI)	32		seIh	
13	A" (ɔ)	8		pɔt	
15	E (ɛ)	4		sɛp	
17	I' (I)	2		tɪt	
19	A (æ)	5		pæs	
20	A'I (aI)	102		kaI	
22	A" (ɔ)	8		tɔt	
22	A" (ɔ)	8		əɔt	
24	I' (I)	2		sɪs	
24	I' (I)	2		tɪt	
25	U" (ʌ)	12		bʌt	
26	EI (eI)	32		keIt	
29	E (ɛ)	4		fɛk	
31	E (ɛ)	4		sɛk	
34	A'I (aI)	102		taIp	

Figure 3.3. Analysis Work Sheet.

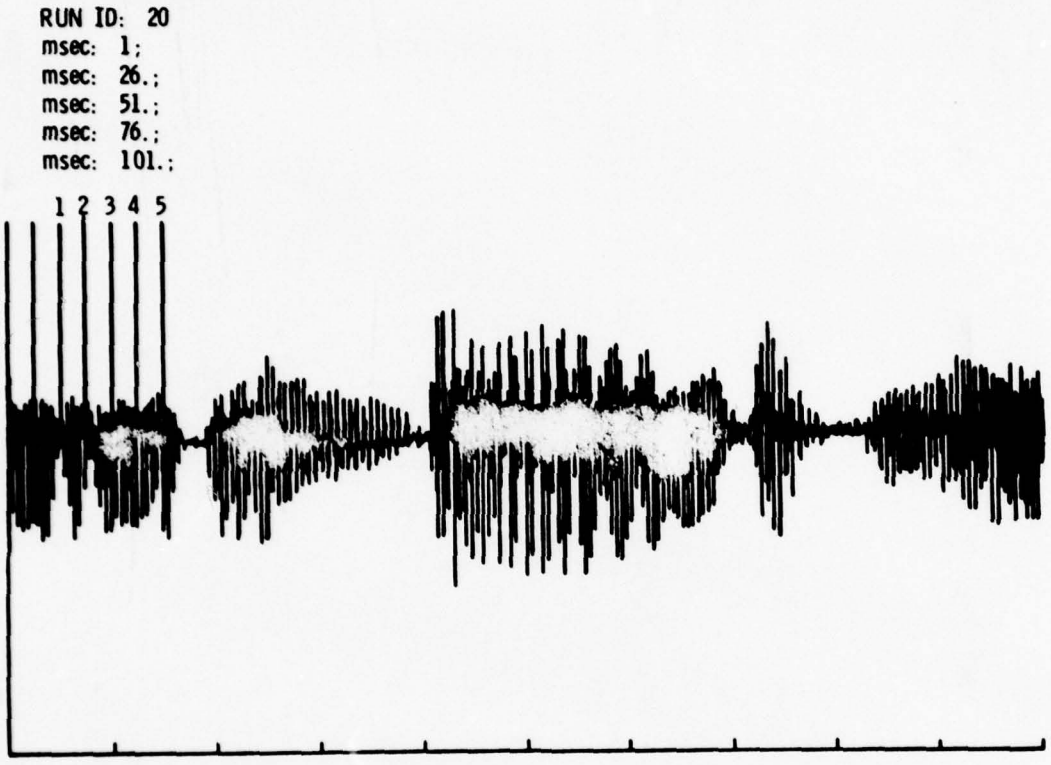

Figure 3.4. EXTRAC Program Output, Form 1.

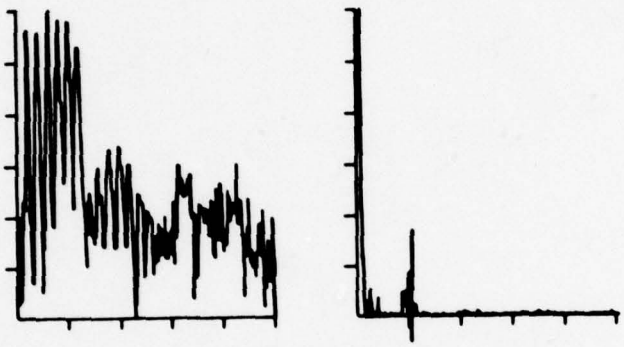

Figure 3.5. EXTRAC Program Output, Form 2.

CEPSTRUM FILE LISTING

COUNT	INDEX	ID	STRT	SPKR	VOWEL
1	2	16	5480	100	2
3	6	18	4410	100	32
5	8	18	8970	100	32
7	10	18	9640	100	32
9	13	21	7760	100	5
11	15	21	8570	100	5
13	21	30	1120	100	8
15	24	32	2710	100	1
17	30	38	2400	100	32
19	33	38	2610	100	32
21	36	39	3800	100	8
23	39	47	6690	100	4
25	43	53	2640	100	97
27	46	60	9110	100	9
29	5	4	40	100	5
31	7	4	550	100	5
33	10	5	9280	100	102
35	14	11	7720	100	8
37	16	11	8390	100	8
39	21	12	1690	100	8
41	25	21	3260	100	12
43	32	36	7840	100	4
45	4	16	30	101	32
47	6	16	3230	101	32
49	8	16	3730	101	32

Figure 3.6. Partial Cepstrum Tape Reference Listing.

```
                    ALL VOWEL 4

GROUP  1      CONTAINS 23
GROUP  2      CONTAINS 43   44
GROUP  3      CONTAINS 76
GROUP  4      CONTAINS 95   96   97   98
GROUP  5      CONTAINS 99   100   101   102   103   104
GROUP  6      CONTAINS 143
GROUP  7      CONTAINS 144
GROUP  8      CONTAINS 173   174   175
GROUP  9      CONTAINS 176
GROUP 10      CONTAINS 213   214   215
GROUP 11      CONTAINS 247   248   249
GROUP 12      CONTAINS 286
GROUP 13      CONTAINS 382
```

Figure 3.7. SPKTST Program Output, Form 1.

```
                                SAMPLE 1

ACCUMULATOR STATUS:      -0.88807     0.40404     -0.19301     0.02161

 AUTO     2      -0.44404          0.09852

 AUTO HIST  0  0  0  0  0  0  0  0  0  1  0  0  1  0  0  0  0  0  0  0

 AUTO HIST  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0

 CROSS    4      -0.04825          0.06402

 CROSS HIST  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2

 CROSS HIST  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

Figure 3.8. SPKTST Program Output, Form 2.

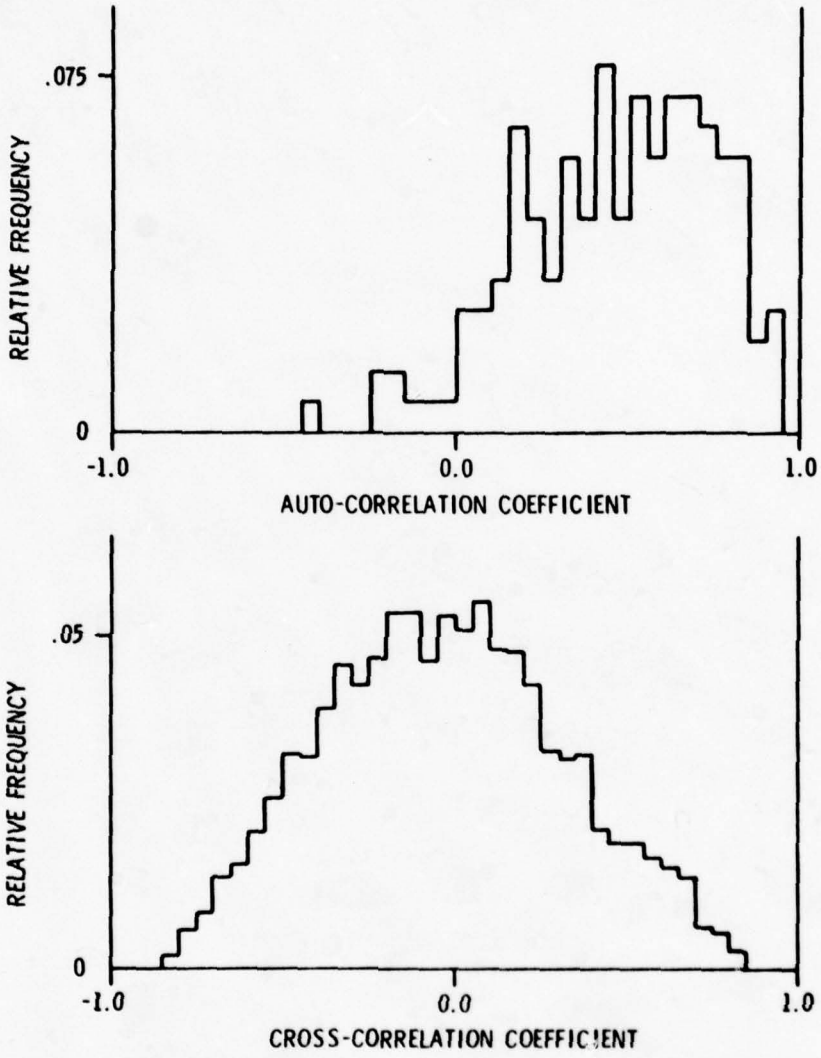

Figure 4.1. Vowel 4 Correlation Coefficient Histograms.

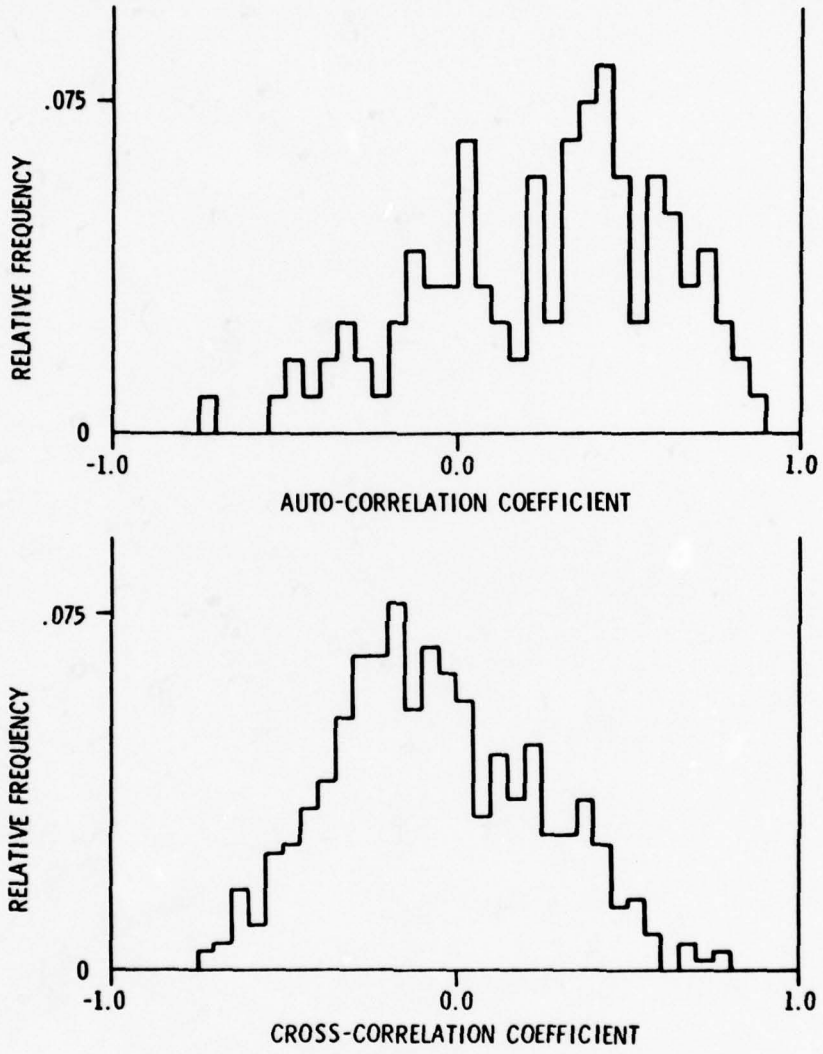

Figure 4.2. Group 1 Correlation Coefficient Histograms.

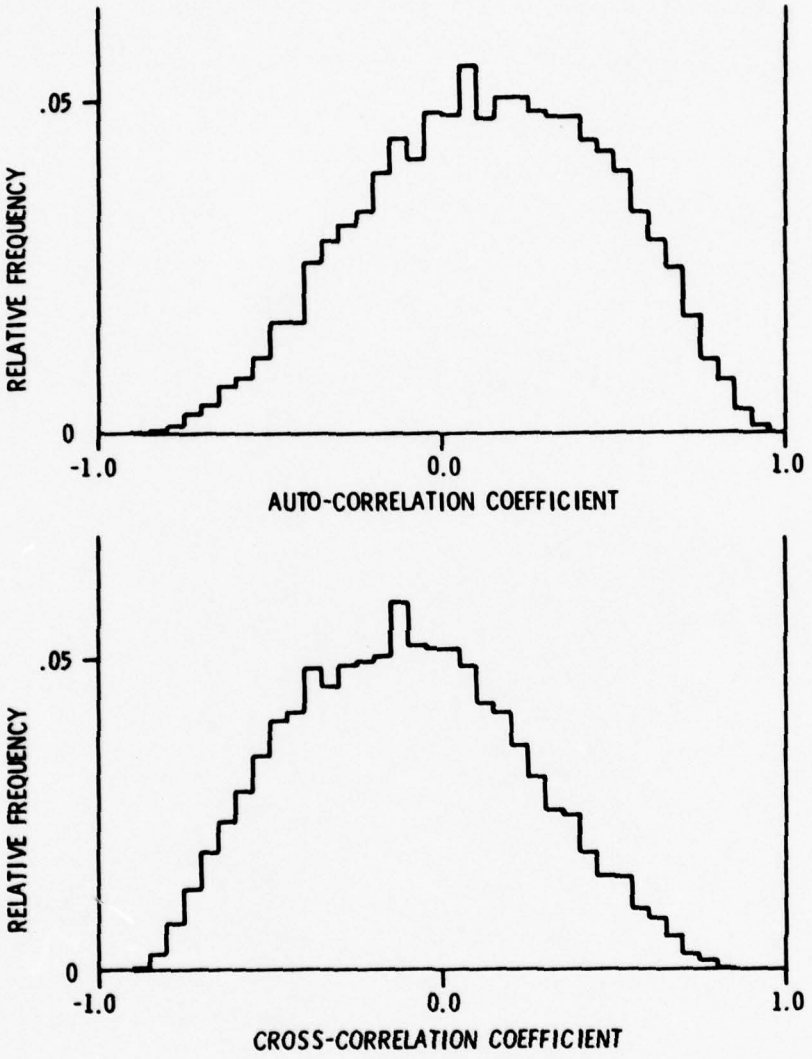

Figure 4.3. Sex Determination Correlation Coefficient Histograms.

TABLE III.1

SUBJECT SPEAKER DATA

SUBJECT	GROUP	AGE	SEX	YEARS SMOKING	SINGER[1]	ALCOHOL[2]
100	1	32	M	3	N	0
101	1	52	M	15	N	0
102	1	55	M	40	N	0
103	1	18	M	0	N	0
104	1	48	M	0	T	0
105	2	28	M	0	S	0
111	2	22	M	8	N	N
120	2	37	M	20	N	0
179	2	23	M	0	T	0
108	3	40	F	20	S	0
228	3	20	F	5	N	F
173	3	25	F	5	N	0
232	3	21	F	4	N	0
238	3	21	F	5	S	F
213	3	20	F	10	N	0
133	4	47	F	0	S	0
162	4	21	F	0	T	0
116	4	18	F	0	N	0
119	4	21	F	0	N	0
263	4	25	F	0	S	0
124	4	23	F	0	S	0

It was originally intended that Group 1 should include male smokers, Group 2 male nonsmokers, Group 3 female smokers, and Group 4 female nonsmokers.

[1]This is the coded response to a question: T-trained, S-singer, N-no.

[2]Use of alcoholic beverages: F-frequent, O-occasional, N-never.

TABLE IV.1

SPEAKER IDENTIFICATION SUMMARY

VOWEL NUMBER	n_a	μ_a	σ_a	n_c	μ_c	σ_c	SPEAKER GROUP
102	9	.193	.317	57	-.095	.255	1
32	65	.263	.307	244	-.048	.290	1
8	40	.279	.350	131	-.122	.286	1
12	20	.047	.436	85	-.065	.290	1
5	7	.244	.450	48	.003	.260	1
4	18	.196	.434	102	-.044	.329	1
2	1	-.647	.000	2	-.357	.100	1
102	5	-.008	.101	16	-.106	.278	2
32	27	.297	.286	93	-.090	.280	2
12	24	.136	.314	54	-.146	.260	2
8	30	.242	.294	61	-.212	.227	2
5	5	.401	.256	23	-.155	.215	2
4	17	.393	.230	38	-.268	.314	2
2	9	-.045	.388	12	-.145	.162	2
102	8	-.137	.334	70	-.000	.360	3
32	33	.341	.280	220	.070	.286	3
12	4	-.215	.454	41	-.076	.349	3
8	36	.161	.287	240	-.029	.337	3
5	6	.476	.138	60	-.057	.342	3
4	18	.439	.230	135	-.090	.319	3
2	2	-.116	.431	26	-.108	.402	3
102	6	.053	.375	60	-.057	.342	4
32	30	.480	.256	201	.011	.377	4
8	30	.184	.357	201	-.036	.298	4
5	8	.433	.249	70	.028	.350	4
4	18	.344	.295	135	-.092	.298	4

TABLE IV.2

SPEAKER IDENTIFICATION GROUPED STATISTICS

IDENT	n_a	μ_a	σ_a	P_a*	n_c	μ_c	σ_c	P_c*
vowel 4	155	.459	.280	.09	3005	-.030	.348	---
vowel 5	59	.396	.270	.65	1481	.013	.329	.03
vowel 8	238	.375	.318	.15	4712	-.023	.350	---
vowel 32	223	.467	.253	---	4727	.019	.349	---
group 1	120	.253	.345	.30	525	-.064	.292	.01
group 2	84	.285	.281	---	215	-.159	.274	.13
group 3	92	.265	.314	.05	656	-.009	.324	---
group 4	86	.325	.333	---	607	-.028	.335	.035

*The value of P_a and P_c is evaluated by determining χ-squared, defined by

$$\chi^2 = \Sigma^{(f-f_c)^2}/f_c$$

and determining the probability that the sample distribution is randomly selected from a normally-distributed universe.

TABLE IV.3

SPEAKER CLASSIFICATION GROUPED STATISTICS

TYPE OF TEST	n_a	μ_a	σ_a	P_a	n_c	μ_c	σ_c	P_c
Sex Determination	7192	.132	.347	---	7408	-.104	.334	---
Female Smoking	1440	.073	.336	---	1537	-.008	.318	---
Male Smoking	396	.052	.309	---	424	-.015	.310	---
Long/Short Lapses	236	.066	.314	.50	272	-.090	.310	.07

TABLE IV.4

χ-SQUARE VALUES

NUMBER	TYPE	χ-SQUARED PROBABILITY
155	auto vowel	.09
59	auto vowel	.65
238	auto vowel	.15
223	auto vowel	---
120	auto group	.30
84	auto group	---
92	auto group	.05
86	auto group	---
7192	auto class	---
1440	auto class	---
396	auto class	---
236	auto class	.50
3005	cross vowel	---
1481	cross vowel	.03
4712	cross vowel	---
4727	cross vowel	---
525	cross group	.01
215	cross group	.13
656	cross group	---
607	cross group	.035
7408	cross class	---
1537	cross class	---
424	cross class	---
272	cross class	.07

"auto" refers to an auto-correlation distribution, "cross" to a cross-correlation distribution. The term "vowel" refers to an accumulated vowel distribution, "group" to an accumulated group distribution, both of which are identification type exercises. The term "class" refers to a speaker classification test, for example, on the basis of sex or smoking.

TABLE V.1

IDENTIFICATION PROBABILITY TABLE

TYPE	P_t	P_f
Vowel 4	.770	.230
Vowel 5	.709	.291
Vowel 8	.723	.277
Vowel 32	.773	.227
Group 1	.682	.318
Group 2	.796	.204
Group 3	.658	.342
Group 4	.686	.314

TABLE V.2

SPEAKER IDENTIFICATION PREDICTED PERFORMANCE

TYPE	$P_t(1)$	$P_f(1)$	$P_t(2)$	$P_f(2)$	$P_i(2)$	$P_t(3)$	$P_f(3)$	$P_i(3)$
Vowel 4	.770	.230	.592	.053	.355	.456	.012	.532
Vowel 5	.709	.291	.503	.085	.412	.357	.025	.618
Vowel 8	.723	.277	.523	.075	.402	.381	.020	.599
Vowel 32	.773	.227	.597	.052	.351	.462	.012	.526
Group 1	.682	.318	.466	.101	.433	.318	.032	.650
Group 2	.796	.204	.633	.042	.325	.504	.009	.487
Group 3	.658	.342	.433	.117	.450	.285	.040	.675
Group 4	.686	.314	.471	.098	.431	.323	.031	.646

TABLE V.3

SPEAKER IDENTIFICATION RELIABILITY TABLE

TYPE	R(1)	R(2)	R(3)
Vowel 4	..770	.918	.974
Vowel 5	.709	.855	.935
Vowel 8	.723	.874	.950
Vowel 32	.773	.920	.975
Group 1	.682	.822	.908
Group 2	.796	.938	.982
Group 3	.658	.787	.877
Group 4	.686	.828	.912

TABLE V.4

SPEAKER CLASSIFICATION PROBABILITY TABLE

TYPE	P_t	P_f
Sex Determination	.633	.367
Female Smoking	.546	.454
Male Smoking	.626	.374
Long/Short Lapse Between Readings	.620	.380

TABLE V.5

SPEAKER CLASSIFICATION PREDICTED PERFORMANCE

TYPE	$P_t(1)$	$P_f(1)$	$P_t(2)$	$P_f(2)$	$P_i(2)$	$P_t(3)$	$P_f(3)$	$P_i(3)$
Sex Det.	.633	.367	.400	.135	.465	.254	.049	.697
F/Smoking	.546	.454	.298	.206	.496	.163	.093	.744
M/Smoking	.626	.374	.391	.140	.469	.245	.052	.703
L/S Lapse	.620	.380	.384	.145	.471	.238	.055	.707

TABLE V.6

SPEAKER CLASSIFICATION RELIABILITY TABLE

TYPE	R(1)	R(2)	R(3)
Sex Det.	.633	.748	.838
F/Smoking	.546	.591	.637
M/Smoking	.626	.736	.825
L/S Lapse	.620	.726	.812

APPENDIX A

THE RAINBOW PASSAGE

This is a standard text passage for phonetic tests. Readings of this passage by a number of speakers constitute the data base for this thesis. Two criteria were applied to select vowels for inclusion in the study: first, that they must be immediately surrounded by voiceless consonants, and second, that they must occur more than once in the passage. The following vowels were selected:

VOWEL	NUMBER	CONTEXT
/I/	2	beautiful
/ei/	32	take
/ei/	32	shape
/æ/	5	path
/ʌ/	12	...according
/ɔ/	8	pot
/ʌ/	12	but
/ei/	32	say he
/ɔ/	8	pot
/ɛ/	4	accepted
/I/	2	that it
/æ/	5	passed
/ɑi/	102	sky...

VOWEL	NUMBER	CONTEXT
/ɔ/	8	Aristotle
/ɔ/	8	thought
/I/	2	physicists
/I/	2	that it
/ʌ/	12	but
/ei/	32	complicated
/ɛ/	4	the effect
/ɛ/	4	the second
/ɑi/	102	common type

THE RAINBOW PASSAGE

When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.

Throughout the centuries, men have explained the rainbow in various ways. Some have accepted it as a miracle without physical explanation. To the Hebrews, it was a token that there would be no more universal floods. The Greeks used to imagine that it was sign from the gods to foretell war or heavy rain. The Norsemen considered the rainbow as a bridge over which the gods passed from the earth to their home in the sky.

Other men have tried to explain the phenomenon physically. Aristotle thought that the rainbow was caused by reflection of the sun's rays by the rain. Since then, physicists have found that it is not reflection, but refraction by the raindrops which causes the rainbow.

Many complicated ideas about the rainbow have been formed. The difference in the rainbow depends considerably upon the size of the water drops, and the width of the colored band increases as the size of the drops increases. The actual primary rainbow observed is said to be the effect of superposition of a number of bows. When the red of the second bow falls upon the green of the first, the result is to give a bow with an abnormally wide yellow band, since red and green lights, when mixed, form yellow. This is a very common type of bow, one showing mainly red and yellow, with little or no green or blue.

APPENDIX B

MEAN VOWEL SPECTRA AND FUNDAMENTAL FREQUENCIES

It is easy to determine the pitch from an examination of the cepstrum. Indeed, one researcher has compared the old view of speech encoding, in which most of the attention is directed to the spectrum determination and the pitch determination is a little black box, to the use of cepstrum pitch determination which produces the spectrum as a by-product!

The fundamental frequency is derived from each of the cepstra for a given vowel and speaker. It is shown as the average value for each of the speaker classes. These numbers are presented in Table B.1.

The part of the cepstrum near the origin may be regarded as an approximation to the vocal tract impulse response. It is therefore possible to derive the vocal tract spectrum from the cepstrum. To obtain the figures presented in this appendix, the following procedure was used.

The complex cepstra from each utterance were summed to give the average cepstrum of each utterance. The average cepstra were reduced to unity mean, and separately accumulated for each class of speakers and over all classes. A cutoff in the cepstrum domain corresponding to a speaker fundamental frequency of 250 Hertz was used so as to preserve as much of the spectrum fine structure as possible while removing the pitch-frequency fluctuations. Vowel spectra are shown for two vowels accumulated for all male and all female speakers. The differences in

the quality of the vowels is quite noticeable in the spectrum, and the difference between male and female utterances shown by the spectra is typical of that observed in all vowels.

The operation of convolution of the glottal pressure wave with the vocal tract impulse response in the time domain, appears in the frequency domain as multiplication of the vocal tract frequency response with the glottal pressure wave spectrum. One may estimate the observed spectrum by multiplying the given vocal tract frequency response by a glottal pressure wave spectrum such as is shown in Figure 2.3.

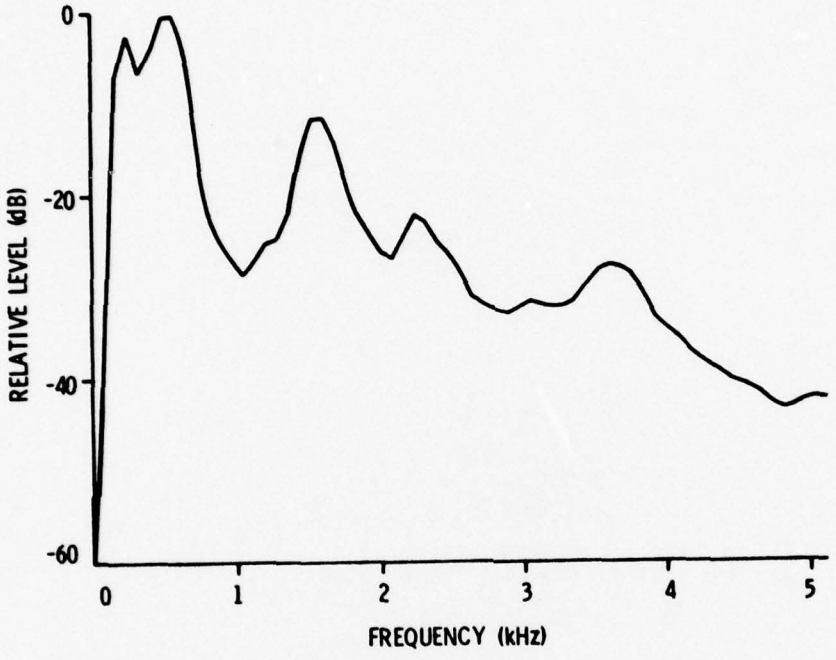

Figure B.1. Vowel 4, Male Speakers.

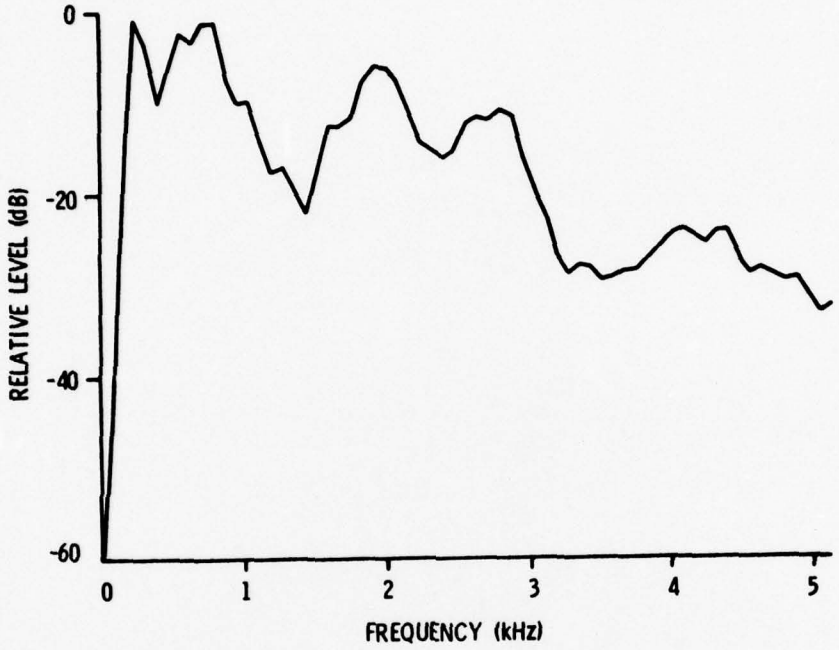

Figure B.2. Vowel 4, Female Speakers.

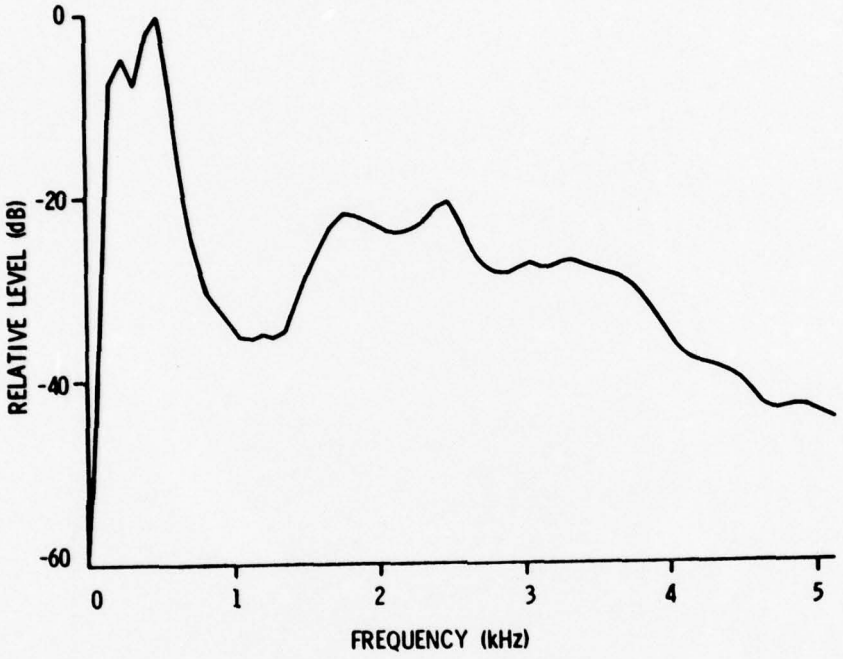

Figure B.3. Vowel 32, Male Speakers.

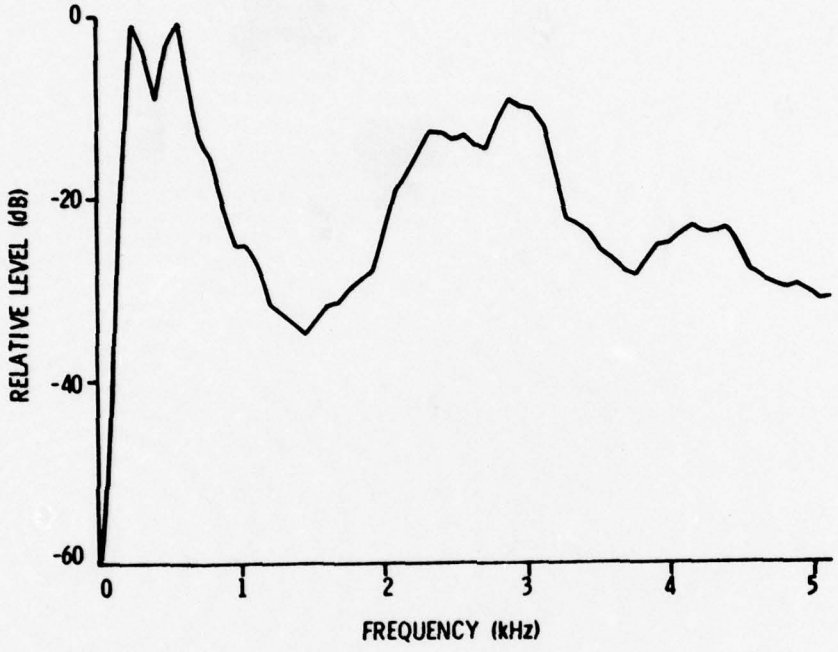

Figure B.4. Vowel 32, Female Speakers.

TABLE B.5

FUNDAMENTAL FREQUENCIES

TYPE	NUMBER	f_0
Vowel 4 Group 1	35	114.2
Vowel 4 Group 2	24	136.8
Vowel 4 Group 3	43	202.4
Vowel 4 Group 4	44	207.8
Vowel 5 Group 1	59	108.0
Vowel 5 Group 2	41	118.2
Vowel 5 Group 3	64	198.7
Vowel 5 Group 4	58	189.5
Vowel 8 Group 1	54	107.1
Vowel 8 Group 2	41	120.8
Vowel 8 Group 3	85	195.3
Vowel 8 Group 4	93	202.1
Vowel 32 Group 1	79	113.2
Vowel 32 Group 2	53	124.6
Vowel 32 Group 3	98	196.0
Vowel 32 Group 4	86	203.7
All Group 1	226	110.5
All Group 2	159	123.6
All Group 3	290	197.3
All Group 4	281	200.7
All Male	385	115.3
All Female	571	199.0

APPENDIX C

FIELD-MODIFIED TELEPHONE BOOTH SOUND ISOLATION CHARACTERISTICS

The sound isolation characteristics of the field-modified telephone booth were measured on April 21, 1976. The booth was installed in the north-west corner of the Hammond Building Museum room at The Pennsylvania State University. An ILG noise source, Serial 17-05-066AS, and a General Radio sound level meter, Model GR1558A S/N 344, were used in the measurements. The physical configuration used in the test was the same as that used during the speech recording session.

The telephone booth is constructed of metal and plexiglass with some internal acoustical damping material. A batten was constructed, consisting of two sections, each four feet by eight feet. These were covered with six inches of fiberglass, and a layer of thin muslin to prevent raveling. A heavy carpet was placed on the floor under the booth, and a cloth drape was used to close the entrance. The physical setup is shown in Figure C.1.

Table C.2 shows the measured octave-band sound pressure levels under various conditions. "Booth" denotes the conditions shown in Figure C.1; "Batten" indicates that the batten and carpet are present but booth is removed; and "Bare Wall" indicates that all of the sound isolation apparatus is removed.

The conclusions of this study indicate that the ILG source is responsible for most of the noise when it is on. The ambient noise was

unmeasurable at 9600-19200 Hz, but this is well above the range of interest. The sound booth is ineffective in the 37.5-150 Hz range, marginal from 150 to 300 Hz, and effective from 300 to 9600 Hz.

AD-A061 858 PENNSYLVANIA STATE UNIV UNIVERSITY PARK APPLIED RESE--ETC F/G 17/2
THE IDENTIFIABILITY OF APPROXIMATE VOCAL TRACT IMPULSE RESPONSE--ETC(U)
DEC 77 F S MCKENDREE N00017-73-C-1418
UNCLASSIFIED ARL/PSU/TM-77-331 NL

2 OF 2
AD
A061858

END
DATE
FILMED
2-79
DDC

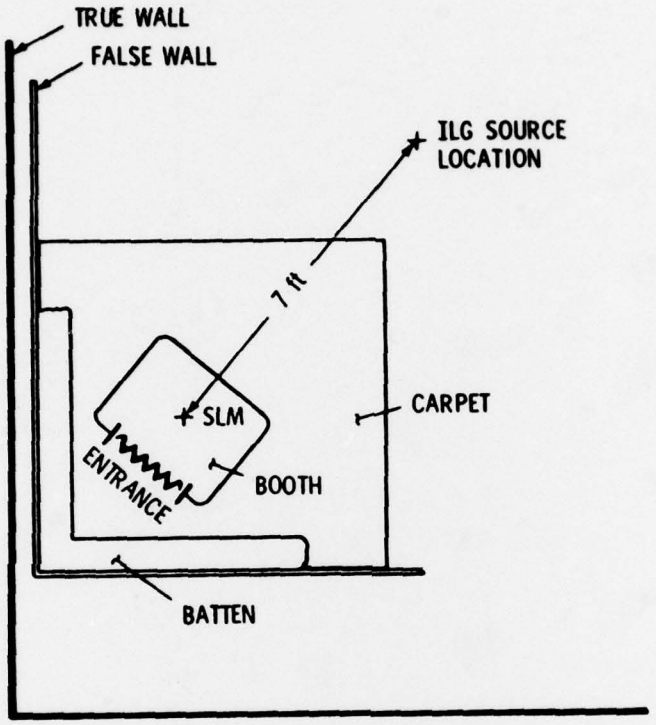

Figure C.1. Modified Telephone Booth Test Setup.

TABLE C.2

SOUND LEVEL MEASUREMENTS

FREQUENCY BAND	----BARE WALL---- ILG off	ILG on	BATTEN ILG on	------BOOTH------ ILG off	ILG on
37.5-75	70	77	76	68	79
75-150	62	68	69	60	67
150-300	52	69	69	52	61
300-600	50	68	69	*	57
600-1200	*	71	70	*	54
1200-2400	*	72	70	*	51
2400-4800	*	70	68	*	49
4800-9600	*	63	63	*	*
9600-19200	*	47	47	*	*

*Sound pressure level less than 46 dB in indicated octave band, which was the lower limit of measurement in this meter.

REFERENCES

CHAPTER I

1. J. C. Steinberg, J. Acoust. Soc. Am. 6, 16-24 (1934).

2. R. K. Potter, Proc. Inst. of Radio Engineers 18, 581-648 (1930).

3. W. Koeing, H. K. Dunn, and L. Y. Lacey, J. Acoust. Soc. Am. 18, 19-49 (1946).

4. L. G. Kersta, Nature 196, 1253-1257 (1962).

5. O. Tosi, H. Oyer, W. Lashbrook, C. Pedrey, J. Nicol, and E. Nash, J. Acoust. Soc. Am. 51, 2030-2043 (1972).

6. O. Tosi, paper presented at XIV International Congress on Logopedics and Phoniatrics, Paris, September 1968.

7. M. A. Young and R. A. Campbell, J. Acoust. Soc. Am. 42, 1250-1254 (1967).

8. K. N. Stevens, C. E. Williams, J. P. Carhonell, and B. Woods, J. Acoust. Soc. Am. 44, 1496-1607 (1968).

9. S. Pruzansky, J. Acoust. Soc. Am. 35, 354-358 (1963).

10. "A Semi-Automatic Speaker Identification System," U.S. Dept. of Justice Grant ni-71-078-g, R. Becker, F. Clarke, F. Poza, and J. Young, October 1973.

11. R. H. Bolt, *et al.*, J. Acoust. Soc. Am. 47, 597-612 (p. 607)(1970).

CHAPTER II

1. R. Timcke, H. von Leden, and P. Moore, Am. Med. Assoc. Arch. Otol. 68, 1-19 and 26-45 (1958).

2. K. N. Stevens, "Acoustical Aspects of Speech Production," Chapter 9 of Handbook of Physiology.

3. P. Liberman, J. Acoust. Soc. Am. 35, 344-353 (1962).

4. K. N. Stevens and A. S. House, J. Acoust. Soc. Am. 27, 484-493 (1955).

5. J. C. Steinberg and N. R. French, J. Acoust. Soc. Am. 18, 4-18 (1946).

6. S. E. G. Ohman, J. Acoust. Soc. Am. 40, 979-988 (1966).

7. S. E. G. Ohman, J. Acoust. Soc. Am. 39, 151-168 (1966).

8. K. N. Stevens and A. S. House, J. Speech and Hearing Res. 6, 111-128 (1963).

BIBLIOGRAPHY

"An Acoustical Theory of Vowel Production and Some of Its Implications," K. N. Stevens and A. S. House, J. Speech and Hearing Res. 4, 75-92 (1961).

Applied General Statistics, Croxton and Cowden (New York: Prentice-Hall, 1939).

"Cepstrum Pitch Determination," A. M. Noll, J. Acoust. Soc. Am. 41, 293-309 (1967).

"On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies," C. G. M. Fant, from Roman Jakobson (The Hague: Mouton, 1956).

"Short-Term Spectrum and 'Cepstrum' Techniques for Vocal Pitch Determination," A. M. Noll, J. Acoust. Soc. Am. 36, 296-302 (1964).

"Speech Analysis/Synthesis System Based on Homomorphic Filtering," A. V. Oppenheim, J. Acoust. Soc. Am. 45, 458-465 (1969).

"Toward the Specification of Speech," R. K. Potter and J. C. Steinberg, J. Acoust. Soc. Am. 22, 807-820 (1950).

The preceding have provided valuable background information but were either not directly quoted in the thesis or so often used that their inclusion in the references was impractical.

DISTRIBUTION

Commander (NSEA 09G32)
Naval Sea Systems Command
Department of the Navy
Washington, DC 20362 — Copies 1 and 2

Commander (NSEA 0342)
Naval Sea Systems Command
Department of the Navy
Washington, DC 20362 — Copies 3 and 4

Defense Documentation Center
5010 Duke Street
Cameron Station
Alexandria, VA 22314 — Copies 5 through 16